연구보고서 2020-10

전국 단위 실태조사 표본설계 효율화 방안 연구

– 장애인실태조사를 중심으로

이혜정
오미애·이민경·손창균·박승환·진재현·염아림

사람을 생각하는 사람들

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

【연구책임자】
이혜정 한국보건사회연구원 부연구위원

【공동연구진】
오미애 한국보건사회연구원 연구위원
이민경 한국보건사회연구원 부연구위원
손창균 동국대학교 교수
박승환 강원대학교 교수
진재현 한국보건사회연구원 전문연구원
염아림 한국보건사회연구원 전문원

연구보고서 2020-10

전국 단위 실태조사 표본설계 효율화 방안 연구
- 장애인실태조사를 중심으로

발 행 일 2020년 12월
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 ㈜삼일기획

ISBN 978-89-6827-724-5 93510

발|간|사

최근 가구특성의 변화와 더불어 예기치 못한 감염병의 확산, 개인정보 보호에 대한 인식의 강화 등으로 대면조사 환경은 악화되고 조사비용은 증가하는 추세에 있다. 이러한 상황에도 불구하고 표본조사를 통해 수집된 자료는 정책을 마련하고 그 효과를 평가하는 데 중요한 기반이 되므로 필수불가결하다고 할 수 있다.

점점 어려워지는 조사환경에 대응하고자 우리 연구원에서 주관하고 있는 장애인실태조사를 중심으로 표본설계 효율화 방안에 대한 기초연구를 살펴보았다. 장애인실태조사는 조사구를 기반으로 표본추출한 조사대상 가구를 조사하며 조사대상은 4만 가구 이상으로 많은 편에 속한다. 조사구를 사용하는 조사의 부담을 줄일 수 있는 방안으로 가구 및 판별조사의 비중을 줄이고, 장애인의 심층조사 확대를 고려해 볼 수 있다. 이 방안과 더불어 현행 조사방법도 다각적으로 검토하였으며 연구 내용은 다음과 같다.

첫 번째는 장애인 출현율을 산출할 수 있는 범위 내에서 표본조사구를 기존보다 축소할 수 있는지에 대한 것이다. 2017년 장애인실태조사 자료를 사용하여 표본조사구 축소 관련 모의실험을 실시하였다. 두 번째는 표본추출틀을 인구센서스 기반 조사구와 등록장애인 DB를 병행할 수 있는지에 대한 것이다. 이와 관련하여 이중추출틀 방법론을 고찰하고 장애인실태조사에서의 활용 방안을 모색하였다. 또한, 장애인 규모 추정 시 기존 방법과 다른 새로운 통계적 방법을 시도하였다.

이 연구는 이혜정 부연구위원의 책임 하에 우리 연구원의 오미애 연구위원, 이민경 부연구위원, 진재현 전문연구원, 염아림 전문원이 연구진으로 참여하였다. 외부 연구진으로 동국대학교 손창균 교수, 강원대학교 박

승환 교수가 참여하였다. 모든 연구진의 노고에 감사드린다. 보고서 작성과 관련하여 유익한 의견을 주신 원내 김성희 연구위원, 원외 청주대학교 류제복 교수, 표본조사 관련 전문가, 그리고 익명의 검독위원들에게도 감사의 마음을 전한다.

마지막으로 이 보고서의 내용은 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝힌다.

2020년 12월

한국보건사회연구원 원장

조 흥 식

목 차

표 목차

그림 목차

Abstract

A Study for Efficient Sample Design: Focusing on the National Survey of Disabled Persons

Project Head: Lee, Hyejung

The current environment, characterized by rapid changes in household conditions, unexpected spread of infectious diseases, and increased awareness of personal information protection, makes it hard and costly to conduct face-to-face surveys. Despite these circumstances, however, in-person survey data remain essential to policymaking.

The National Survey of Disabled Persons covers more than 40,000 households. There are two ways to reduce the burden on a survey based on enumeration districts. The first is to reduce the sample size of household and discrimination surveys. The second is to expand the survey of the disabled based on data other than enumeration districts.

The existing and proposed methods were examined in various ways. The existing method is a survey based on enumeration districts. The proposed method is a survey based on both enumeration districts and databases of individuals registered as disabled. The target population for both methods is composed of disabled persons.

Co-Researchers: Oh, Miae· Yi, Mingyeong· Son, Changkyoon· Park, Seunghwan· Jin Jaehyun· Yeom, Ahrim

The existing method consists of a single frame and a double sampling technique. The advantage of the existing method is that it can calculate the prevalence of individuals with disabilities and produce reliable data. The disadvantage of this method is that it is costly and comes with the burden of having to cover an increasing number of households. A further study is needed to determine the effective sample size for the first phase.

The proposed method employs a dual frame and a stratified multistage sampling technique. The method excels at calculating the prevalence of individuals with disabilities even with a limited budget and enables appearance rate of the disabled under the available budget and enables conduct additional surveys if necessary. The disadvantage of this method is that it requires more skilled manpower and time compared to the existing method, and the difficulty with survey management. From the findings from our simulation, we propose that the enumeration districts can be reduced by 30% compared to the National Survey of Disabled Persons (2017). A further study needs to be conducted on the effective survey sample size for unregistered disabled persons.

We examined the population size of the disabled using the capture-recapture method. The results show that the proposed estimator is similar to the known population size (National Survey of Disabled Persons, 2017). The proposed estimator has a considerably smaller standard error.

* Key words: National Survey of Disabled Persons, sample design, dual frame, Capture-Recapture method

요약

1. 연구의 배경 및 목적

최근 1인 가구 증가와 같은 가구특성의 변화와 더불어 예기치 못한 감염병 확산, 개인정보 보호에 대한 인식 강화 등으로 기존의 대면조사 환경은 악화되고, 조사비용은 증가하는 추세에 있다. 이러한 상황에도 불구하고 표본조사를 통해 수집된 자료는 정책을 마련하고, 그 효과를 평가하는 데 중요한 기반이 되므로 필수불가결하다고 할 수 있다.

이렇듯 점점 어려워지는 조사환경에 대응하기 위해서는 표본설계의 효율화 방안에 대한 기초연구가 필요한 상황이다. 이에 본 연구는 우리 연구원에서 주관하고 있는 장애인실태조사를 중심으로 방안을 살펴보고자 한다.

장애인실태조사는 조사구를 기반으로 표본추출한 가구를 대상으로 조사하였으며, 조사대상은 4만 가구 이상으로 꽤 많은 편에 속한다. 조사구를 사용하는 조사에 대한 부담을 줄일 수 있는 방안으로 「2020년 장애인실태조사 사전연구」에서 2가지를 제안하였다(김성희 외, 2019, pp.74~75). 첫 번째 방안은 현행 유지 방안이고, 두 번째 방안은 가구 및 판별조사의 비중을 줄이고 장애인에 대한 심층조사를 확대하는 방안이다.

본 연구의 목적은 2가지 방안을 다각적으로 살펴보는 것이며 연구 내용은 다음과 같다. 첫 번째, 장애인 출현율을 산출할 수 있는 범위 내에서 표본조사구를 기존보다 축소할 수 있는지에 대한 것이다. 장애인 출현율 제시는 장애인실태조사의 목적 중 하나이므로 조사방식이 변경되더라도 산출될 수 있어야 한다. 따라서 장애인 출현율 조사를 위한 가구조사 비중을 줄일 수 있는지에 대한 모의실험을 실시해 본다. 이때 조사구 축소 비율은 다양하게 설정하여 살펴본다. 두 번째, 표본추출틀을 인구센서스

와 더불어 등록장애인 DB도 활용할 수 있는지에 대한 것이다. 즉, 등록장애인 DB를 활용한 리스트 조사 병행 가능성을 검토하여 등록장애인에 대한 심층조사 확대 가능성을 가늠해 보고자 한다.

2. 주요 연구결과

장애인실태조사 표본설계 효율화 방안으로 다음과 같이 2가지를 검토하였다(〈요약표 1-1〉 참조). 목표모집단은 모두 장애인이고 각 방안별 특징을 살펴보면 다음과 같다.

가. 조사구를 활용하는 방안(현행 방안)

첫 번째는 조사구를 활용하는 방안으로 현재 장애인실태조사에서 실시하고 있으며, 현행 유지 방안이라고 볼 수 있다. 조사모집단은 등록센서스 기반 조사구로 단일추출틀이고, 표본추출은 이중추출방법을 이용한다. 이 방안의 장점은 기존 방식으로 조사를 실시하므로 장애인 출현율을 산출할 수 있을 뿐만 아니라 이전 조사와의 종적 비교도 가능하므로 통계량의 신뢰성도 상대적으로 높은 편이다. 실사 관리에 있어서도 그동안의 노하우(knowhow)로 돌발 상황 등에도 유연하게 대처할 수 있기에 안정적이라고 볼 수 있다. 그러나 다음과 같은 3가지 우려 사항이 발생할 수 있다. 첫 번째, 가구 및 판별조사를 대규모로 실시하기 때문에 조사비용이 많이 소요된다는 점이다. 두 번째, 해가 거듭될수록 조사환경 변화로 인하여 3만 가구 이상을 조사 완료해야 하는 부담감이 있다는 점이다. 세 번째, 물가상승에 따른 조사 제반 비용(인건비, 조사수수료 등)의 인상분이 예산에 반영되지 않는다면 향후 예산 부족 문제로 이어질 수 있다는 점이다.

향후 보완 방안으로 1상에서의 표본 규모 관련 비용과 시간을 절약하기 위한 최적값 산출, 등록장애인 DB를 활용한 표본추출틀 재구축과 관련한 심층연구가 필요하다고 생각한다.

나. 조사구와 등록장애인 DB를 병행하는 방안

두 번째는 조사구와 등록장애인 DB를 병행하는 방안으로 가구 및 판별조사의 비중은 줄이고 장애인에 대한 심층조사를 확대하는 것이다. 조사모집단은 인구센서스를 바탕으로 작성된 조사구 및 등록장애인 DB로 이중추출틀이고, 표본추출은 층화다단추출방법을 이용한다.

이 방안의 장점은 인구센서스를 바탕으로 작성된 조사구 조사를 실시하므로 적정한 조사구 규모 하에서 장애인 출현율 산출이 가능하다. 또한 등록장애인 DB도 함께 활용하므로 장애 관련 정보 변수(장애유형, 장애등급 등)를 고려한 표본 배분이 가능하여 희귀질환 장애유형도 표본에서 누락되지 않고 구축할 수 있다. 덧붙여 조사구 조사의 규모가 축소되어 가용할 수 있는 조사예산이 확보되므로 추가로 심층조사를 실시할 수 있다. 그러나 예상되는 우려 사항은 첫 번째, 2개의 추출틀 간 기준시점의 차이로 표본 포함 범위가 다르게 되는 상황이 발생할 수 있다는 점이다. 두 번째, 등록장애인 DB는 동일한 지역으로 묶고 집락을 만들어서 관리를 하는데 리스트이므로 표본관리가 어려울 수 있다는 점이다. 세 번째, 이중추출틀을 활용한 조사는 기존 조사에 비해 표본추출에서부터 가중치 산출 관련 통계 업무 및 실사관리가 기존 조사에 비해 더 많은 연구인력이 필요하고 투입해야 할 시간도 더 많이 소요된다는 점이다.

한정된 조사예산이므로 조사구와 등록장애인 DB에서의 표본 배분을 검토해야 한다. 그래서 다양한 표본조사구 축소 비율에 따른 최대허용오

차 한계를 살펴보고, 축소된 조사구 조사로 인해 가용할 수 있는 조사예산으로 추가 심층조사가 가능한 등록장애인 규모도 가늠해 보았다.

먼저 다양한 표본조사구 축소 비율에 따른 모의실험 결과는 다음과 같다. 조사의 공표 범위인 권역별(대도시, 중소도시, 농어촌)에 대한 단순임의추출 가정으로 95% 신뢰수준 하에서 가구수 기준 최대허용오차 한계는, 표본조사구 축소 30% 이하의 경우 모두 1% 내외의 값을 가졌다. 반면에 표본조사구 축소 40% 이상에서는 모두 1% 이상으로 나타났다. 한편 조사구 축소 비율에 따른 전체 장애인 규모는 표본조사구를 10% 축소한 경우 6,121명으로 산출되었으며 모집단(6,820명) 대비 89.8%를 차지하였다. 즉, 표본조사구를 10% 축소하면 조사대상자 규모도 약 10% 축소된다고 볼 수 있다. 표본조사구를 20% 축소한 경우에는 5,460명(80.1%), 30%의 경우 4,777명(70.0%), 40%의 경우 4,098명(60.1%)이고 50%의 경우 3,416명(50.1%)으로 나타났다(〈요약표 1-1〉 참조). 모의실험 결과를 종합해 보면, 표본조사구를 축소한 비율만큼 조사대상 규모도 비슷한 비율로 축소되는 양상을 보였다. 가구수 기준 최대허용오차의 한계 및 조사대상 규모의 결과를 통하여 기존 표본조사구 규모의 30% 이하까지는 축소 가능하다고 볼 수 있다. 단, 조사의 정확도 및 조사비용을 종합적으로 고려한 최대허용오차에 대한 기준을 마련하고 조사구 조사를 통해서만 가능한 미등록 장애인의 구축 방안 등에 대한 심층연구를 실시해야 할 필요가 있다.

현재 실시 중에 있는 국내 장애인 관련 실태조사 예산 집행을 기반으로 하여, 2017년 장애인실태조사 예산 기준으로 심층조사가 추가 가능한 표본 수를 산출해 보았다. 단, 조사 사례비 수준에 따라 예산 변동이 가능하고, 조사구와 등록장애인 DB에서 추출된 표본의 중복 가능성을 반영해야 하는 등과 같은 여러 가지 제약 조건이 있다. 이러한 제약 조약 하에서 절

감된 조사예산에 맞춰 대략적으로 추가할 수 있는 심층조사 가능한 장애인의 규모를 추정해 보았다. 〈요약표 1-1〉을 보면 예상되는 전체 장애인 수는 표본조사구 축소 10%인 경우 6,821~7,121명, 20%인 경우 7,260~7,660명, 30%인 경우 7,777~8,277명, 40%인 경우 8,098~8,598명, 50%인 경우 8,416~8,916명으로 나타났다. 표본조사구 축소 비율이 증가할수록 2017년 실태조사 응답 완료 기준 6,594명보다 더 많은 장애인을 조사할 수 있다. 그러나 조사구를 활용한 조사의 표본 규모가 축소되면 미등록 장애인을 구축하는 데 어려움이 발생할 수 있다. 이에 따라 충분하지 않은 미등록 장애인의 규모는 대표성 문제를 야기할 수 있고 가중치 산출에도 어려움이 따를 수 있다. 표본조사구 축소 시 미등록 장애인의 특성을 파악하여 분포가 많은 곳은 과대표본추출을 고려해 볼 수 있다.

부가적인 연구로 장애인 규모 추정에 대해 Capture-Recapture 방법(이하 C-R 방법)을 사용하여 추정해 보았다. C-R 방법으로 장애인구 수를 추정한 결과는 기존(2017년 장애인실태조사에서의 추정 장애인구 수)과 근소한 차이를 보였다. 표준오차의 값은 C-R 방법에서 기존보다 작아지는 것을 확인하였다. 또한 장애유형, 성·연령, 층화변수(대도시/중소도시/농어촌)를 기준으로 각각을 구분하여 장애인구 수도 추정해 보았다. 장애유형, 성·연령, 층화변수 모두 C-R 방법이 기존보다 장애인 규모를 적게 추정하는 것으로 나타났다. 특히 층화변수를 사용했을 때 C-R 방법에 따른 추정 장애인구 수가 장애유형, 성·연령을 기준으로 추정했을 때보다 적게 추정되었다.

〈요약표 1-1〉 조사구를 활용하는 방안 vs. 조사구와 등록장애인 DB를 병행하는 방안

	조사구를 활용하는 방안	조사구와 등록장애인 DB를 병행하는 방안
목표모집단	장애인	장애인
조사모집단 (표본추출틀)	등록센서스 기반 조사구 (단일추출틀)	등록센서스 기반 조사구 및 등록장애인 DB (이중추출틀)
표본추출방법	이중추출방법	층화다단추출방법
장점	· 장애인 출현율 산출 가능 · 이전 조사와의 종적 비교 가능하여 통계량의 신뢰성 높은 편 · 실사 관리 측면에서 보면 축적된 경험으로 인한 유연한 상황 대처 가능	· 적정한 조사구 규모 하에서 장애인 출현율 산출 가능 · 조사구 조사 축소로 인한 가용할 수 있는 조사예산 확보로, 추가 심층조사 가능 · 등록장애인 DB의 활용으로 희귀질환 장애유형도 표본에서 누락되지 않고 구축 가능
단점(우려)	· 대규모 가구 및 판별조사로 인한 많은 조사비용 소요 · 해가 거듭될수록 조사환경 변화로 인하여 3만 가구 이상 조사 완료해야 하는 부담감 · 조사 제반 비용의 인상분이 예산에 반영되지 않는다면 향후 예산 부족 발생	· 2개 추출틀 간 기준시점 차이로 표본 포함 범위가 다르게 되는 상황 발생 · 표본관리 어려움 · 기존 조사에 비해 더 많은 연구인력 필요 및 더 많은 투입 시간 소요
보완 방안	· 1상에서 표본 규모 관련 최적값 산출을 위한 심층연구 필요 · 표본설계 고도화를 위한 등록장애인 DB 활용 방안 마련(표본추출틀 재구축 등)	· 표본조사구 축소 시 미등록 장애인 구축 방안 마련(과대표본추출 등)

	2017년 조사 기준	표본조사구 축소				
		10%	20%	30%	40%	50%
조사구 (개)	2,001	1,798	1,602	1,400	1,203	1,001
가구수 (가구)	36,200	32,520	28,987	25,328	21,767	18,095
A : 전체 장애인 수 (명)	6,820	6,121	5,460	4,777	4,098	3,416
B : 추가 심층조사 가능한 등록장애인 수 (명)	-	700~ 1,000	1,800~ 2,200	3,000~ 3,500	4,000~ 4,500	5,000~ 5,500
A+B (추정 규모) : 전체 장애인 수 (명)	-	6,821~ 7,121	7,260~ 7,660	7,777~ 8,277	8,098~ 8,598	8,416~ 8,916

주: 전체 예산 중 직접비를 70%로 가정하였을 때 표본조사구 축소 10%의 경우 조사비용은 7% 절감으로 추정됨. 표본조사구 축소 20%인 경우 조사비용은 14%, 30%인 경우 21%, 40%인 경우 28%, 50%인 경우 35%임.

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출 및 저자 작성.

3. 결론

이 연구는 장애인실태조사의 표본설계 효율화를 위하여 기존 및 새로운 방안에 대해 다각적으로 검토한 기초연구라고 볼 수 있다. 다만 2017년 장애인실태조사 자료 분석에 한정하였고, 조사차수별 분석 및 추이를 살펴볼 수 없었던 제약으로 인하여 통계적 분석 결과는 일반화할 수 없다는 점을 밝혀 둔다.

현재 실시하고 있는 조사방법은 여러 가지 상황에 비추어 볼 때 적합한 방법이라고 볼 수 있다. 그러나 등록장애인 DB 정보의 활용과 관련해서는 계속 제고해야 하고, 실제 조사에 바로 적용하기보다는 심층연구를 통한 신중한 접근이 필요하다고 생각한다.

* 주요 용어: 장애인실태조사, 표본설계 효율화 방안, 이중추출틀, Capture-Recapture 방법

제1장

서론

제 1 장 서론

제1절 연구 배경 및 목적

최근 1인 가구 증가와 같은 가구특성의 변화와 더불어 예기치 못한 감염병의 확산, 개인정보 보호에 대한 인식 강화 등으로 기존 대면조사 환경은 악화되고 있으며, 조사비용은 증가 추세에 있다. 이러한 상황에도 불구하고 표본조사(sample survey)를 통해 수집된 자료(data)는 정책을 마련하고 그 효과를 평가하는 데 중요한 기반이 되므로 필수불가결하다고 할 수 있다.

보통 국내 표본조사는 인구센서스 기반, 즉 조사구를 사용하여 조사하고 있다. 예를 들면 장애인실태조사, 한국복지패널조사, 한국의료패널조사, 재정패널조사, 한국노동패널조사 등 다수의 국내 표본조사들이 이에 해당된다. 한편, 인구센서스를 기반으로 표본을 추출할 경우 조사 시점의 간극으로 추출한 조사구와 현재 조사구 간 불일치 가능성이 존재한다. 예를 들면 우리 연구원에서 주관하고 있는 장애인실태조사는 2017년에 조사를 실시하였는데 이때 2015년 등록센서스 기반 조사구를 사용하였다. 이러한 시점 차이는 재개발, 건물용도 변경, 신도시 개발 등과 같은 변화가 조사구 변동으로 이어지게 된 것이다. 조사구 변동은 조사구를 탐색하는 데 있어 애로사항으로 볼 수 있으며 더 나아가 조사비용의 상승 요인이 될 수 있다. 이렇듯 점점 어려워지는 조사환경에 대응하고자 표본설계에 대한 효율화 방안을 마련하기 위한 기초연구가 필요한 상황이다. 이 연구는 장애인실태조사를 중심으로 살펴보고자 한다.

장애인실태조사는 첫 조사가 1980년에 시작된 이후 1985년, 1990년, 1995년, 2000년, 2005년, 2008년, 2011년, 2014년, 2017년에 조사를 실시하였다. 2005년 이후로는 조사 주기가 3년이며, 올해(2020년) 제11차 조사를 실시하고 있다. 올해는 코로나19로 인한 조사 여건 등의 어려움으로 인해 등록장애인 DB를 기반으로 조사대상자를 선정하였다. 2008년 비장애인에 대해 조사를 실시한 것과 관련한 지적이 있어서 등록장애인 DB를 기반으로 장애인을 조사한 경험이 있다. 나머지 조사 연도에서는 인구센서스를 기반으로 하였다.

장애인실태조사의 목적은 2가지이다. 첫 번째 목적은 우리나라 장애인구를 파악하고, 우리나라 유일의 장애인 출현율을 산출하는 것이다(김성희 외, 2018, p.33). 두 번째는 장애인에 대한 심층조사를 실시하여 장애인의 생활실태 및 복지욕구를 파악하고 복지서비스 이용 등에 대한 신뢰성 있는 데이터를 생산하며, 장단기 장애인복지정책 수립과 추진을 위한 기초 자료를 제공하는 것이다(김성희 외, 2018, p.33).

조사표는 가구 및 판별조사표, 심층조사표로 구성되어 있으며, 조사 실시 원칙은 2가지이다. 첫째, 가구 및 판별조사표에서 법정장애인이 없는 가구로 판정된 가구에 대해서는 가구 및 판별조사표만 작성한다. 둘째, 가구 및 판별조사에서 법정장애인이 있는 가구로 1차 판정된 가구에 대해서는 가구 및 판별조사표의 작성과 더불어 장애인 심층조사표를 모두 작성한다(김성희 외, 2018, p.80).

장애인실태조사는 조사구를 기반으로 표본추출한 가구를 조사하며 조사대상은 4만 가구 이상으로 꽤 많은 편에 속한다. 이렇듯 조사구를 사용하는 조사에 대한 부담을 줄일 수 있는 방안으로 「2020년 장애인실태조사 사전연구」에서 다음과 같이 2가지를 제안하였다(김성희 외, 2019, pp.74~75). 첫 번째 방안은 현행 유지 방안으로 전년도 결과와의 종적

비교가 가능하다는 장점이 있으나, 조사환경의 어려움 등이 우려된다고 할 수 있다(〈표 1-1〉 참조).

〈표 1-1〉 현행 유지 방안

대안	장점	단점 (우려)
○ 출현율을 위한 가구 조사 실시 ○ 가구조사 시 발견된 장애인에 대한 심층조사 실시	○ 장애인실태조사의 전년도 결과와 종적 비교 가능 (출현율, 장애인 실태 현황)	○ 조사환경의 어려움 등 고려 필요 - 표본틀은 조사시점의 t-2 인구센서스 활용 • 시기적 차이가 있고, 신도시의 경우 조사 어려움 - 예산 부족의 고려 • 조사원 인건비 상승 등을 반영하기 어려움

자료: 김성희 외(2019).

두 번째 방안은 가구 및 판별조사의 비중을 줄이고 장애인에 대한 심층조사를 확대하는 방안이다(〈표 1-2〉 참조). 즉, 표본조사구 수를 감축하는 대신 다른 정보를 활용하여 표본을 추출하는 조사와 병행하는 것이다. 여기서 활용할 수 있는 다른 정보로는 등록장애인 DB가 있다. 등록장애인 DB는 등록장애인에 대한 현황을 파악한 자료로, 효율적인 정책수립 및 지원을 위해 작성된 보고통계 자료이다(보건복지부, 2018, p3). 이 방안의 장점은 장애인 심층조사의 확대에 있다. 그러나 우려 사항으로 장애인 출현율의 신뢰성에 대한 위험 가능성이 있으므로 통계청 승인을 위한 면밀한 분석이 필요하다. 특히 조사구를 사용한 조사와 등록장애인 DB 등을 통한 리스트 조사를 병행하는 부분에 대한 검토가 필요하다.

〈표 1-2〉 출현율 제시 유지 및 장애인에 대한 심층조사 확대 방안

대안	장점	단점 (우려)
○ 출현율 조사를 위한 가구조사 비중 축소 ○ 장애인 심층조사-리스트 조사 병행 (장애인 심층조사의 동반 감소를 보완, 강화 위함) ※ 관련 전문가의 자문 필요 - 출현율 조사를 위한 가구조사 비중을 줄이는 부분 등에 대한 전문가 검토 및 논의 필요	○ 장애인 심층조사의 확대	○ 출현율 신뢰성에 위험 가능성 - 가구조사의 조사구, 조사 가구수를 줄이면 출현율을 위한 장애인 수 추정에 무리가 있을 수 있음 ○ 통계청 승인을 위한 면밀한 분석 필요 - 조사구를 사용한 조사와 등록장애인 DB 등을 통한 리스트 조사를 병행하는 부분에 대한 면밀한 분석 필요

자료: 김성희 외(2019).

이 연구의 목적은 2가지 방안을 다각적으로 살펴보는 것이며 연구 내용은 다음과 같다. 첫 번째, 장애인 출현율을 산출할 수 있는 범위 내에서 표본조사구를 기존보다 축소할 수 있는지에 대한 것이다. 장애인 출현율 제시는 장애인실태조사의 목적 중 하나이므로 조사방식이 변경되더라도 산출될 수 있어야 한다. 그래서 장애인 출현율 조사를 위한 가구 및 판별조사의 비중을 줄일 수 있는지에 대한 모의실험을 실시해 볼 것이며, 이때 조사구 축소 비율은 다양하게 설정하여 살펴본다. 두 번째, 표본추출틀을 인구센서스와 더불어 등록장애인 DB도 활용할 수 있는지에 대한 것이다. 등록장애인 DB를 활용한 리스트 조사 병행 가능성을 검토하여 등록장애인에 대한 심층조사 확대 가능성을 가늠해 보고자 한다.

제2절 연구 내용 및 방법

1. 연구 내용

이 연구는 전국 단위 실태조사의 표본설계 효율화 방안 연구로 장애인 실태조사를 중심으로 살펴본다. 각 장에서의 연구 내용은 아래와 같다.

제2장은 1995년부터 최근 조사를 완료한 2017년까지의 장애인실태조사의 모집단, 층화 및 표본배분 등과 관련한 표본설계 현황을 살펴보고, 장애인 출현율 정의 및 현황을 파악해 본다. 또한 2017년 실태조사 자료에 대한 표본조사구 분포, 가구 및 가구원 특성 분석을 실시한다.

제3장은 표본조사구 축소 관련 모의실험, 이중추출틀을 활용한 표본추출 방법론, 통계적 추정 방법에 대해 살펴본다. 먼저 표본조사구 축소 관련 모의실험은 2017년 장애인실태조사 자료에 근거하여 실시하며 목적은 다양한 표본조사구 축소 비율에 따른 가구수 변화 및 효과를 살펴보는 것이다. 다양한 상황을 고려해 볼 수 있도록 최저 10%에서부터 최대 50%까지 극대화(10%, 20%, 30%, 40%, 50%)하여 실시하며, 각 상황에서의 장애인 규모에 대한 변화를 살펴본다. 평가지표는 단순임의추출 가정으로 95% 신뢰수준 하에서 가구수 기준 최대허용오차 한계이다.

다음은 이중추출틀을 활용한 표본추출 방법론으로 동일한 모집단에서 2개의 추출틀을 사용하는 방법론을 소개하고 장·단점도 살펴본다. 그리고 이중추출틀을 활용하여 표본설계한 해외 사례도 소개하고, 한정된 조사 자원 하에서 이중추출틀을 활용할 때 최적으로 할당하기 위한 고려사항 및 통계적 방법도 검토하고자 한다.

마지막 통계적 추정 방법은 주로 생태학 분야에서 개체군의 규모를 추정하기 위해 사용하는 방법 중 하나인 Capture-Recapture 방법론이다.

최근에는 생태학 분야 이외에도 보건학 분야에서 특정 질병 총 환자 수, 마약 사용자 수, 도시의 노숙자 인원수 등과 같이 인구 모집단을 추정하는 데도 많이 활용되고 있다. 따라서 2017년 장애인실태조사 자료에 Capture-Recapture 방법론을 사용하여 전체 장애인 규모를 추정해 보고, 기존의 추정 규모와 비교해 본다.

제4장은 조사구를 활용하는 방안(현행 방안)과 조사구와 등록장애인 DB를 병행하는 방안에 대해 다각적으로 검토하고 향후 과제를 제안한다.

2. 연구 방법

그동안 실시한 장애인실태조사의 표본설계 및 장애인 출현율 현황을 파악해 본다. 이를 비롯하여 국내외 선행연구, 자료 분석, 전문가 자문회의 등을 활용하여 장애인실태조사의 표본설계에 관한 효율화 방안을 면밀히 검토해 본다. 표본조사구 축소관련 모의실험, 통계적 추정 방법에서 사용한 자료는 2017년 장애인실태조사 자료이고, 통계패키지는 SAS를 사용한다.

제2장

장애인실태조사 표본설계 고찰

제 2 장 장애인실태조사 표본설계 고찰

제1절 표본설계 현황

장애인실태조사의 표본설계 현황을 장애인실태조사 보고서를 바탕으로 하여 작성하였는데 1995년부터 2017년까지의 조사 연도별 모집단, 층화 및 표본배분을 비교하였다.

목표모집단은 장애인으로 정의한다. 장애인은 「장애인복지법」 제2조(장애인의 정의 등)에 따라 '신체적, 정신적 장애로 오랫동안 일상생활이나 사회생활에서 상당한 제약을 받는 자'로 정의한다. 여기서 '신체적 장애'는 주요 외부 신체기능의 장애, 내부기관의 장애 등을 말하는 것이고, '정신적 장애'는 발달장애 또는 정신질환으로 인해 발생하는 장애를 말한다. 본 조사에서의 장애인은 「장애인복지법」 제32조(장애인 등록)에 따라 장애인등록을 한 자(이하 등록 장애인)로 국한하지 않고, 장애인등록을 하지 않았으나 장애상태에 있는 자(이하 미등록 장애인)를 모두 조사 대상에 포함한다.

한편 이러한 장애인 중 시설에 수용되어 있는 경우는 1980년 조사 당시에는 별도로 조사를 수행하지 않았고, 1985년에는 보건사회부의 기존 통계자료를 인용하였다(김성희 외, 2018, p.35). 1990년부터 2008년까지는 사회복지시설에 수용되어 있는 장애인을 대상으로 우편설문조사를 실시하였다(1990년 677개소, 1995년 748개소, 2000년 875개소, 2005년 1,063개소, 2008년 1,068개소). 그리고 2011년부터는 시설에 거주하는 장애인은 조사에서 배제하였다(김성희 외, 2018, pp.35~36).

조사모집단(표본추출틀)을 살펴보면, 1995년에는 1990년도 인구총조사의 조사구, 2000년에는 1995년 인구주택총조사의 조사구, 2005년에는 2000년 인구주택총조사 조사구를 이용하여 대상지역을 선정하였다(〈표 2-1〉 참조). 반면 2008년에는 비장애인을 조사하는 것에 대한 지적이 있어서 등록장애인만을 대상으로 조사를 실시하였고, 등록장애인 DB를 사용하여 조사대상자를 선정하였다. 2011년에는 다시 2005년도 인구주택총조사 보통조사구를, 2014년에는 2010년 인구주택총조사 보통조사구를, 2017년에는 2015년 등록센서스 기반 표본추출틀의 조사구를 이용하였다.

2008년을 제외하고 표본추출틀로 사용한 인구주택총조사(등록센서스) 자료를 통해 장애인 가구의 판별은 불가능하므로 조사모집단과 목표모집단 간 차이가 발생할 수 있다(김성희 외, 2011, p.52, 김성희 외, 2014, p.56, 김성희 외, 2018, p.68). 이와 같이 표본추출틀의 과소 포괄성 문제가 있지만, 현실적으로 장애인 모집단 자료는 인구주택총조사 자료를 사용할 수밖에 없기에 추가적으로 장애인 시설에 대한 조사를 통해 보완할 수 있을 것이다(김성희 외, 2011, p.52, 김성희 외, 2014, p.56, 김성희 외, 2018, p.68).

〈표 2-1〉 조사모집단

연도	정의
2017년	2015년 등록센서스 기반 표본추출틀의 조사구에 거주하는 일반가구 및 가구원 (기숙사, 특수 시설 조사구, 섬지역 조사구 제외)
2014년	2010년 인구주택총조사 90% 조사구 중 보통조사구
2011년	2005년 인구주택총조사 90% 조사구 중 보통조사구
2008년	2007년 11월 보건복지부 등록장애인 DB에 기재된 장애인 명부
2005년	2000년 인구주택총조사 조사구 자료 2000년 인구주택총조사 이후의 신축 아파트 자료
2000년	1995년 인구주택총조사 조사구 자료 신축아파트 자료
1995년	1990년 인구총조사 조사구 자료

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

장애인실태조사의 표본설계는 표본에 15개 장애유형이 모두 포함되어 장애유형별 복지서비스 욕구가 적절히 반영되어야 한다(김성희 외, 2011, p.53, 김성희 외, 2014, p.57, 김성희 외, 2018, p.69). 표본추출틀은 이러한 장애유형에 대한 정보가 없기 때문에 표본설계에서는 반영할 수 없으며, 조사모집단과 목표모집단 간에도 가구와 개인으로 추정 수준이 다른 한계점도 있다(김성희 외, 2011, p.53, 김성희 외, 2014, p.57, 김성희 외, 2018, p.69). 표본배분 단계에서 이러한 점을 보완하기 위하여 제한적인 정보이지만, 장애인과 장애인 가구는 서로 연관성이 높으므로 지역별 장애인 분포를 이용하였다(김성희 외, 2011, p.53, 김성희 외, 2014, p.57, 김성희 외, 2018, p.69).

〈표 2-2〉를 보면 2017년과 2014년은 대도시, 중소도시, 농어촌의 3개 층, 2011년은 4개 층, 2005년은 6개 층, 2000년은 5개 층, 1995년은 3개 층으로 층화하였다. 이는 시도별로 장애인 출현율이 낮은 특정 장애인을 포괄할 수 있도록 하기 위해서였다.

한편 2008년에는 전국 16개 시도별 읍면동을 집락 및 층화(3,573개 읍면동)하여 전국 16개 시도별 읍면동을 1차 추출단위로 하였고, 추출된 표본 읍면동(208개)의 등록장애인을 지역별 표본 읍면동의 15개 장애유형과 6개 장애등급으로 층화한 후 2차 추출단위인 등록장애인을 추출하였다(변용찬 외, 2009, pp.64~65).

〈표 2-2〉 층화 기준

연도	정의
2017년	대도시(서울, 부산, 대구, 인천, 광주, 대전, 울산의 동부), 중소도시(경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 세종시의 동부), 농어촌(9개 도 및 세종시 읍·면 지역)의 3개 층
2014년	대도시(서울, 부산, 대구, 인천, 광주, 대전, 울산의 동부), 중소도시(경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주의 동부), 농어촌(9개 도 읍·면 지역)의 3개 층
2011년	서울 동부, 광역시 동부, 중소시도의 동부, 전국 읍·면·부의 4개 층
2008년*	읍면동(집락), 전국 16개 시·도(층화)
2005년	서울의 동, 광역시의 동, 도의 시의 동, 광역시의 읍·면, 도의 시의 읍·면, 도의 군의 읍·면의 6개 층
2000년	서울시의 동, 광역시(부산, 대구, 인천, 광주, 대전, 울산)의 동, 기타 시의 동, 읍, 면의 5개 층
1995년	6대 시, 기타 시, 군부의 3개 층

주: 2008년은 1차 추출단위에 대한 정보임.
자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

〈표 2-3〉을 보면 조사구 수는 2011년 이후 층별 등록장애인 규모를 고려하여 1,000개 수준을 추출하였고, 장애출현율을 고려하여 조사구 당 45가구를 조사하도록 하였다. 2005년 이전에는 200개 이하로 나타났는데 2005년의 경우 통합조사구를 표본조사구로 사용하였기 때문이다. 통합조사구란 표본으로 지정된 조사구가 2000년 인구주택총조사의 조사구

인 경우에는, 이 조사구부터 차례로 하나씩 조사구를 추가하여 그 누적 가구수가 220가구를 초과할 때까지의 3~5개의 조사구를 통합한 것이다(변용찬 외, 2006, p.89). 그래서 187개 표본조사구에 포함된 인구주택총조사 조사구는 776개 조사구로 나타났다(변용찬 외, 2006, p.90). 2000년은 조사구당 평균 221가구였으며(변용찬 외, 2000, p.71), 1995년은 200개의 표본조사구에 대해 전수조사(41,283가구 대상)를 실시하였다(정기원, 권선진, 계훈방, 1995, p.66).

가구 및 판별조사의 조사완료 가구수는 2017년 36,200가구로 꾸준히 감소하고 있는 것을 알 수 있다. 2005년에는 40,556가구로 가장 많은 가구를 구축하였다. 조사완료율을 보면 모든 조사연도에서 80% 이상으로 조사 성공률이 높은 편이라고 볼 수 있으며 특히 1995년에 94.6%로 가장 높게 나타났다.

〈표 2-3〉 조사구 수, 대상 가구수, 조사완료 가구수 및 완료율

연도	조사구 수	대상 가구수	조사완료 가구수	완료율
2017년	1,000개	44,161가구	36,200가구	81.9%
2014년	1,004개	48,344가구	38,560가구	79.8%
2011년	1,000개	47,458가구	38,231가구	80.6%
2005년	187개	45,285가구	40,556가구	89.6%
2000년	200개	44,128가구	39,411가구	89.3%
1995년	200개	41,283가구	39,078가구	94.6%

주: 2008년은 표본추출틀이 등록장애인 DB이므로 제외함.
자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

표본추출방법은 등록장애인 DB를 사용하여 표본추출한 2008년을 제외하고는 유사하다고 볼 수 있다(〈표 2-4〉 참조).

〈표 2-4〉 표본추출방법

연도	표본추출방법
2017년	확률 비례 계통 추출
2014년	층화 확률 비례 추출
2011년	층화 확률 비례 추출
2008년	층화집락추출
2005년	확률 비례 계통 추출
2000년	확률 비례 계통 추출
1995년	층화 확률 비례 추출

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

〈표 2-5〉는 추정가구수 대비 표본가구수 비중으로 1995년 0.31%에서 2017년 0.18%으로 꾸준히 감소하고 있다. 조사가 진행될수록 추정가구수는 증가하는 반면에 표본가구수는 비슷하거나 감소하는 상황이라고 볼 수 있다.

〈표 2-5〉 추정가구수 대비 표본가구수 비중

(단위: 가구, %)

연도	추정가구수(A)	표본가구수(B)	(B/A)×100
2017년	19,284,671	36,200	0.18
2014년	18,206,328	38,560	0.21
2011년	17,574,018	38,231	0.22
2005년	15,864,809	40,556	0.26
2000년	14,677,637	39,411	0.27
1995년	12,745,280	39,078	0.31

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

〈표 2-6〉은 장애인 심층조사의 표본규모이다. 조사완료 장애인 수는 1995년에 3,335명이었으나 조사차수가 지남에 따라 증가하여 2017년에는 6,549명이었다. 가구당 가구원수는 1995년은 3.34명/가구로 가장 높고 최근 조사로 갈수록 감소하여 2017년에는 2.53명으로 나타났다.

〈표 2-6〉 조사완료 가구 및 장애인 수

연도	조사완료 가구수	가구당 가구원수	조사완료 가구원수	조사완료 장애인수
2017년	36,200가구	2.53명/가구	91,405명	6,549명
2014년	38,560가구	2.72명/가구	104,703명	6,824명
2011년	38,231가구	2.76명/가구	105,496명	6,010명
2005년	40,556가구	2.94명/가구	119,306명	5,466명
2000년	39,411가구	3.14명/가구	123,721명	4,125명
1995년	39,078가구	3.34명/가구	130,556명	3,335명

주: 2008년 장애인실태조사는 등록장애인 DB로부터 확률비례추출된 14,000명의 장애인을 선정하여 7,000명을 대상으로 조사를 수행.

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

제2절 장애인 출현율 현황

장애인실태조사의 목적 중 하나인 장애인 출현율의 정의 및 현황에 대해 살펴보았다.

장애인 출현율은 우리나라 장애인 규모를 추정하는 수치로 산출식은 다음과 같다.

$$\text{장애인 출현율} = \frac{\text{전체 장애인수}}{\text{장래추계인구수}} \times 100$$

여기서, 전체 장애인수는 재가장애인수와 시설장애인수의 총합이다.

장애인실태조사의 조사대상은 재가장애인으로, 등록 장애인뿐만 아니라 미등록 장애인도 포함한다. 미등록 장애인은 장애인실태조사를 통해서만 현황파악이 가능하므로 이를 통하여 미등록 장애인 규모를 추정할 수 있다.

"인구센서스의 경우 전 국민을 대상으로 하고 있지만 장애인을 파악하기 어렵기 때문에 과소추정의 오류를 범할 수 있을 뿐만 아니라 조사 항목도 매우 제한적일 수밖에 없다. 현 실정에서 가장 바람직한 방법은 전체 가구를 모집단으로 하여 가구표본을 추출하고, 이들 가구의 전체 가구원을 대상으로 장애여부를 판정하는 것이다(김성희 외, 2018, p34)."

이렇듯 우리나라 장애인 추정수를 산출할 때 미등록 장애인의 규모를 포함하므로 미등록 장애인을 조사하는 것은 매우 중요하다고 볼 수 있다.

〈표 2-7〉은 조사연도별 장애인 규모 및 출현율을 보여주고 있다. 2017년 전체 장애인은 2,668,411명(재가장애인: 2,580,340명, 시설장애인: 88,071명)으로 1995년(전체 장애인: 1,053,468명) 이후로 계속 증가하는 추세이다. 그러나 전체 장애인 출현율은 2017년에 5.39%로

2011년(5.61%) 이후로 조금씩 감소하고 있다. 1995년의 경우에는 2.35%로 가장 낮은 장애인 출현율을 가지고 있었다. 전체 장애인 중에서 재가장애인이 시설장애인에 비해 높은 비중을 차지하고 있다. 그래서 재가장애인 출현율과 전체 장애인 출현율 간 차이가 크지 않다. 2017년은 재가장애인 출현율이 5.21%이고 전체 장애인 출현율은 5.39%로 나타났다. 한편 2008년은 비장애인을 조사하는 것에 대한 지적으로 등록 장애인만을 대상으로 조사하여 장애인 출현율을 산출하지 않았다. 그러나 비장애인과 장애인의 복지 수요와 관련해서는 동일한 자료로 비교 분석이 불가능하고, 장애인 출현율에 대해 추정 시 장애가구 및 장애인 발생률의 추계를 할 수 없는 문제점이 있었다(김성희 외, 2018, p.67). 그래서 2011년 실태조사부터 다시 장애인과 비장애인 가구를 모두 조사하였다(김성희 외, 2018, p.67).

조사연도별 장애유형별 장애인 출현율 변화 추이는 〈표 2-8〉과 같다. 장애유형은 지체, 뇌병변, 시각, 청각, 언어, 지적, 자폐성, 정신, 신장, 심장, 호흡기, 간, 안면, 장루요루, 뇌전증 장애이며 총15개로 구분한다.

모든 조사연도별에서 지체장애 출현율이 가장 높으며 다음으로 뇌병변장애, 시각장애, 청각장애 출현율이 높은 편에 속한다. 한편 중복장애 출현율의 경우 2005년, 2000년, 1995년에는 지체장애 출현율 다음으로 높은 것으로 나타났으며 나머지 장애유형 출현율은 높지 않은 편이었다. 가장 최근 조사인 2017년은 지체장애 출현율이 2.51%로 가장 높았으며 뇌병변장애 0.52%, 청각장애 0.52%, 시각장애 0.51% 순으로 나타났다. 중복장애 출현율도 0.42%로 장애유형별 장애인 출현율에서 높은 편에 해당했다. 그 외에 심장 장애와 안면장애 각각 0.01%, 호흡기장애 0.02%로 나머지 장애유형 출현율은 낮은 편으로 나타났다.

〈표 2-7〉 조사연도별 장애인 규모 및 출현율

(단위: 명)

조사연도 (년)	장애인 규모			재가장애인 출현율(%)	전체 장애인 출현율(%)	인구수
	재가장애인	시설장애인	전체			
2017	2,580,340	88,071	2,668,411	5.21	5.39	49,482,802
2014	2,646,064	80,846	2,726,910	5.43	5.59	48,759,677
2011	2,611,126	72,351	2,683,477	5.47	5.61	47,850,677
2008*	2,137,222			-	-	
2005	2,101,057	47,629	2,148,686	4.50	4.59	46,811,593
2000	1,398,177	51,319	1,449,496	2.98	3.09	46,853,554
1995	1,028,837	24,631	1,053,468	2.37	2.35	44,850,801

주: 1) 인구수는 2017년의 경우 2016년 장래인구추계(군인, 등록외국인, 국내거소 신고자 제외), 2014년의 경우 2013년 장래인구추계(군인, 외국인 제외), 2011년의 경우 2010년 인구주택총조사(군인, 외국인 제외), 2005년의 경우 2005년 인구주택총조사 가집계 결과자료(군인, 외국인 제외), 2000년의 경우 2000년 추계인구 잠정집계(사회복지시설 인구 제외)이며, 1995년의 경우 장래인구추계(1995년 7월 1일 기준) 값임.

2) * 2008년 조사는 등록 장애인 규모를 나타냄.

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 변용찬 외(2009), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

〈표 2-8〉 조사연도별 장애유형별 장애인 출현율 현황

(단위: %, 명)

구분	2017년 2)		2014년 2)		2011년 2)3)		2005년 1)		2000년 1)		1995년	
	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수
전체	5.39	2,668,411	5.59	2,726,910	5.61	2,683,477	4.59	2,148,686	3.09	1,449,496	2.37	1,028,837
지체장애	2.51	1,242,785	2.71	1,319,132	2.72	1,303,032	1.99	933,553	1.19	556,861	1.40	608,760
뇌병변장애	0.52	258,121	0.48	234,675	0.59	280,180	0.32	150,756	0.23	109,866	-	-
시각장애	0.51	252,046	0.53	257,492	0.51	245,917	0.42	198,456	0.35	163,309	0.13	57,541
청각장애	0.52	256,018	0.50	245,935	0.50	240,695	0.40	185,911	0.23	109,503	0.26	111,461
언어장애	0.03	15,790	0.03	15,252	0.04	17,010	0.02	10,538	0.03	12,956	0.05	22,264
지적장애	0.38	187,300	0.36	173,296	0.28	131,648	0.12	56,268	0.12	57,780	0.07	32,069
자폐성장애	0.03	13,215	0.02	10,572	0.01	5,880	0.01	3,212	0.01	4,626	-	-
정신장애	0.22	111,031	0.23	112,632	0.23	109,817	0.18	82,492	0.14	64,953	-	-

구분	2017년 2)		2014년 2)		2011년 2)3)		2005년 1)		2000년 1)		1995년	
	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수	출현율	추정수
신장장애	0.15	72,722	0.12	60,790	0.10	48,741	0.06	29,720	0.05	21,685	-	-
심장장애	0.01	6,176	0.02	8,331	0.04	17,852	0.08	35,184	0.08	36,221	-	-
호흡기장애	0.02	11,485	0.03	14,965	0.04	17,068	0.05	23,484	-	-	-	-
간장애	0.02	10,609	0.02	11,162	0.02	8,314	0.02	9,975	-	-	-	-
안면장애	0.01	3,073	0.01	2,702	0.00	2,111	0.01	3,223	-	-	-	-
장루요루 장애	0.03	14,309	0.03	14,833	0.03	14,096	0.03	12,614	-	-	-	-
뇌전증장애	0.02	8,299	0.01	6,610	0.02	9,895	0.02	11,235	-	-	-	-
중복장애	0.42	205,431	0.49	238,532	0.48	231,222	0.86	402,065	0.66	311,736	0.45	196,742

주: 1) 시설장애인 포함.
2) 2011, 2014년의 경우 재가장애인은 중복장애로 별도 산정하였으나, 시설장애인은 행복e음을 통해 파악하여 중복장애에 대한 정보가 없어 한 가지 장애만 가지고 있는 것으로 추정함.
3) 시설 거주 장애인 중 중복 장애를 반영함.

자료: 정기원, 권선진, 계훈방(1995), 변용찬 외(2000), 변용찬 외(2006), 김성희 외(2011), 김성희 외(2014), 김성희 외(2018).

제3절 2017년 조사 결과 현황

3장의 표본설계 효율화를 위한 다각적 접근에 대해 살펴보기 전에 2017년 실태조사 자료를 사용하여 기초분석을 하였다. 가구 특성 분석은 장애인 유무 가구에 따라, 개인 특성 분석은 장애등록여부에 따라 구분한 다음 분석을 실시하였다.

1. 가구 특성 분석

장애인실태조사에 참여한 36,200가구의 주요 특성 변수들 간 연관성을 살펴봄으로써 가구에 장애인이 있는지의 여부에 중요하게 영향을 미치는 특성 변수를 파악하고자 하였다.

〈표 2-9〉는 장애인의 거주지 시도에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 전체 17.2% 가구에 장애인이 있다고 나타났고, 그중 충북 지역이 25.4%로 가장 높게 나타났으며, 광주 지역이 11.8%로 가장 낮게 나타났다. 17개 시도별 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의하다고 볼 수 있으나 뚜렷한 관계를 찾아보긴 힘들다.

〈표 2-9〉 시도에 따른 장애인 유무 가구 (17개 시도)

(단위: 가구)

시도 \ 장애인 유무	없음		있음		전체
	빈도	행%	빈도	행%	
서울	4,270	86.8%	648	13.2%	4,918
부산	1,937	83.6%	379	16.4%	2,316
대구	1,238	85.6%	209	14.4%	1,447
인천	1,732	85.5%	294	14.5%	2,026

장애인 유무 / 시도	없음		있음		전체
	빈도	행%	빈도	행%	
광주	893	88.2%	120	11.8%	1,013
대전	1,215	84.0%	232	16.0%	1,447
울산	240	82.8%	50	17.2%	290
세종	248	86.1%	40	13.9%	288
경기	5,360	85.9%	881	14.1%	6,241
강원	1,437	82.3%	308	17.7%	1,745
충북	1,290	74.6%	439	25.4%	1,729
충남	1,542	81.9%	340	18.1%	1,882
전북	1,019	78.6%	278	21.4%	1,297
전남	2,422	76.2%	756	23.8%	3,178
경북	2,479	81.7%	557	18.3%	3,036
경남	2,051	78.4%	566	21.6%	2,617
제주	589	80.7%	141	19.3%	730
전체	29,962	82.8%	6,238	17.2%	36,200
참고: χ^2 = 377.128, P-value=0.000					

자료: 비공개 자료에 따른 출처 생략.

〈표 2-10〉은 장애인의 거주지를 3개 권역(서울, 광역, 세종 / 중소도시 / 농어촌)으로 나누어 3개 권역에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 농어촌이 22.6%로 가장 높게 나타났고 서울, 광역, 세종이 14.3%로 가장 낮게 나타났다. 3개 권역별 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-10〉 3개 권역에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 3개 권역	없음		있음		전체
	빈도	행%	빈도	행%	
서울, 광역, 세종	11,773	85.7%	1,972	14.3%	13,745
중소도시	11,999	83.0%	2,456	17.0%	14,455
농어촌	6,190	77.4%	1,810	22.6%	8,000
전체	29,962	82.8%	6,238	17.2%	36,200
참고: χ^2 = 243.937, P-value=0.000					

자료: 비공개 자료에 따른 출처 생략.

〈표 2-11〉은 장애인의 거주지를 동읍면으로 나누어 동읍면에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 읍면부가 21.8%로 동부의 15.1% 보다 더 높게 나타났다. 동읍면별 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-11〉 동읍면에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 동읍면	없음		있음		전체
	빈도	행%	빈도	행%	
동부	21,033	84.9%	3,742	15.1%	24,775
읍면부	8,929	78.2%	2,496	21.8%	11,425
전체	29,962	82.8%	6,238	17.2%	36,200
참고: χ^2 = 249.261, P-value=0.000					

자료: 비공개 자료에 따른 출처 생략.

〈표 2-12〉는 장애인의 거주지를 3개 권역 및 동읍면으로 나누어 3개 권역 및 동읍면에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 농어촌 및 읍면부 지역이 22.6%로 가장 높게 나타났고 서울, 광역, 세종 및 동부 지역이 14.3%로 가장 낮게 나타났다. 3개 권역 및 동읍면별 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-12〉 3개 권역 및 동읍면에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 3개 권역 및 동읍면	없음		있음		전체
	빈도	행%	빈도	행%	
서울, 광역, 세종&동부	11,656	85.7%	1,945	14.3%	13,601
서울, 광역, 세종&읍면부	117	81.3%	27	18.8%	144
중소도시&동부	9,377	83.9%	1,797	16.1%	11,174
중소도시&읍면부	2,622	79.9%	659	20.1%	3,281
농어촌&읍면부	6,190	77.4%	1,810	22.6%	8,000
전체	29,962	82.8%	6,238	17.2%	36,200

참고: χ^2 = 274.415, P-value=0.000

자료: 비공개 자료에 따른 출처 생략.

〈표 2-13〉은 대체조사구 여부에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 원조사구가 17.6%로 대체조사구 14.9%보다 더 높게 나타났다. 대체조사구 여부에 따른 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-13〉 대체조사구 여부에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 조사구 여부	없음		있음		전체
	빈도	행%	빈도	행%	
원조사구	25,352	82.4%	5,431	17.6%	30,783
대체조사구	4,610	85.1%	807	14.9%	5,417
전체	29,962	82.8%	6,238	17.2%	36,200

참고: χ^2 = 24.341, P-value=0.000

자료: 비공개 자료에 따른 출처 생략.

〈표 2-14〉는 주택 형태에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 기타(비닐하우스 등)가 가장 높으나 빈도수가 8가구로 매우 작다. 그 다음으로는 단독주택이 21.4%로 가장 높게 나타났고, 아파트 오피스텔이 14.1%로 가장 낮게 나타났다. 주택 형태에 따른 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의하다고 할 수 있다.

〈표 2-14〉 주택 형태에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 주택 형태	없음		있음		전체
	빈도	행%	빈도	행%	
단독주택	11,692	78.2%	3,184	21.4%	14,876
아파트 오피스텔	15,246	85.9%	2,500	14.1%	17,746
연립주택	825	85.0%	146	15.0%	971
다세대주택	1,969	85.1%	345	14.9%	2,314
비거주용 건물내 주택	223	80.2%	55	19.8%	278
기타(비닐하우스 등)	6	42.9%	8	57.1%	14
전체	29,961	82.8%	6,238	17.2%	36,119

참고: χ^2 = 333.455, P-value=0.000

자료: 비공개 자료에 따른 출처 생략.

〈표 2-15〉는 본인을 포함한 총 가구원 수에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 총 가구원 수가 6명 이상이라는 응답이 35.7%로 가장 높게 나타났고, 4명이라는 응답은 10.8%로 가장 작게 나타났다. 본인을 포함한 총 가구원 수에 따른 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-15〉 본인을 포함한 총 가구원 수에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 가구원 수	없음		있음		전체
	빈도	행%	빈도	행%	
1명	7,320	83.7%	1,426	16.3%	8,746
2명	8,896	77.8%	2,539	22.2%	11,435
3명	5,958	85.0%	1,055	15.0%	7,013
4명	6,098	89.2%	740	10.8%	6,838
5명	1,406	81.5%	320	18.5%	1,726
6명 이상	284	64.3%	158	35.7%	442
전체	29,962	82.8%	6,238	17.2%	36,200

참고: χ^2 = 532.304, P-value=0.000

자료: 비공개 자료에 따른 출처 생략.

〈표 2-16〉은 장애인 가구의 월평균 총 가구소득을 4분위로 나누어 총 가구소득에 따른 장애인 유무 가구 현황을 나타내고 있다. 장애인이 있는 가구의 비율을 살펴보면 1분위가 25.8%로 가장 높게 나타났고, 4분위가 12.0%로 가장 작게 나타났다. 월평균 총 가구소득에 따른 장애인이 있는 가구의 차이는 유의수준 5% 하에서 통계적으로 유의한 것으로 나타났다.

〈표 2-16〉 월평균 총 가구소득에 따른 장애인 유무 가구

(단위: 가구)

장애인 유무 / 가구소득	없음		있음		전체
	빈도	행%	빈도	행%	
1분위	6,937	74.2%	2,413	25.8%	9,350
2분위	8,158	82.6%	1,717	17.4%	9,875
3분위	8,198	87.3%	1,195	12.7%	9,393
4분위	6,669	88.0%	913	12.0%	7,582
전체	29,962	82.8%	6,238	17.2%	36,200
참고: χ^2 = 759.410, P-value=0.000					

자료: 비공개 자료에 따른 출처 생략.

〈표 2-17〉은 가구 특성 분석을 통해 유의하다고 나타난 설명변수를 사용한 포화 모형에서, 유의하지 않은 설명변수를 하나씩 제거하여 적합한 로지스틱 회귀분석의 최종 모형에 대한 결과이다. 모형은 3개 권역 및 동읍면, 월평균 총 가구소득, 본인을 포함한 총 가구원 수, 주택형태의 4개 변수를 설명변수로, 종속변수는 장애인 유무 가구로 하여 로지스틱 회귀모형을 적합하였다. 각 변수별 계수값, 표준오차 및 유의확률이 다음과 같다.

〈표 2-17〉 로지스틱 회귀분석 결과

설명변수	최종 모형		
	계수값	표준오차	유의확률
서울, 광역, 세종&동부			.000
서울, 광역, 세종&읍면부	.097	.217	.000
중소도시&동부	.152	.036	.654
중소도시&읍면부	.262	.051	.000
농어촌&읍면부	.324	.038	.000

설명변수	최종 모형		
	계수값	표준오차	유의확률
월평균 총 가구소득	-.002	.000	.000
본인을 포함한 총 가구원수	.171	.014	.000
단독주택			.000
아파트 오피스텔	-.283	.033	.000
연립주택	-.308	.093	.000
다세대주택	-.283	.064	.001
비거주용 건물내 주택	.071	.153	.000
기타(비닐하우스 등)	1.364	.542	.642
상수항	-1.426	.040	.012

자료: 비공개 자료에 따른 출처 생략.

지금까지 장애인실태조사에 참여한 36,200가구에 대한 주요 특성 변수들 간의 연관성을 살펴보았다. 가구 거주지를 17개 시도, 3개 권역, 동읍면으로 나눈 설명변수들과 대체조사구 여부, 주택 형태, 본인을 포함한 총 가구원 수, 월평균 총 가구소득의 설명변수들이 유의수준 5% 하에서 장애인 유무 가구에 따라 통계적으로 유의하게 나타났다.

원조사구이면서 농어촌 및 읍면부의 단독주택에 거주하며 본인을 포함한 총 가구원 수가 6명 이상이고, 월평균 총 가구소득이 1분위인 가구에 장애인이 있을 비율이 높게 나타났다.

또한 대체조사구이면서 서울, 광역, 세종 및 동부의 아파트 오피스텔에 거주하며 본인을 포함한 총 가구원 수가 4명이고 월평균 총 가구소득이 4분위인 가구 중에 장애인이 있을 비율이 낮게 나타났음을 확인할 수 있었다.

2. 가구원 특성 분석

장애인실태조사에 참여한 가구원 6,549명에 대한 주요 특성 변수들 간의 연관성을 살펴봄으로써 장애등록여부에 중요한 영향을 미치는 특성 변수를 파악하고자 하였다. 장애등록여부에 대한 집단 구분은 정형화된 것은 아니지만, 3장의 연구를 위한 기초 통계 분석이라는 점을 밝혀 둔다.

〈표 2-18〉은 장애인의 성별에 따른 장애등록여부 현황을 나타내고 있다. 장애등록을 하지 않은 비율을 살펴보면 여성이 3.3%로 남성 1.6%보다 더 높은 편이다. 성별에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-18〉 성별에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 성별	예		아니오		전체
	빈도	행%	빈도	행%	
남성	3,608	98.4%	58	1.6%	3,666
여성	2,789	96.7%	94	3.3%	2,883
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 20.053, P-value=0.000

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-19〉는 장애인 연령대에 따른 장애등록여부 현황이다. 장애등록을 하지 않은 비율을 살펴보면 10대의 비율이 4%로 가장 높으며 다음으로 70대 이상(3.3%)이 높게 나타났다. 반면 20대의 비율이 0.9%로 가장 낮은 편이었다. 연령대에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-19〉 연령대에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 연령대	예		아니오		전체
	빈도	행%	빈도	행%	
10대	194	96.0%	8	4.0%	202
20대	215	99.1%	2	0.9%	217
30대	275	98.9%	3	1.1%	278
40대	609	97.9%	13	2.1%	622
50대	1,197	98.4%	19	1.6%	1,216
60대	1,436	98.5%	22	1.5%	1,458
70대 이상	2,471	96.7%	85	3.3%	2,556
전체	6,397	97.7%	152	2.3%	6,549
참고: χ^2 = 25.010, P-value=0.000					

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-20〉은 장애인의 최종 학력에 따른 장애등록여부 현황에 대한 것이다. 장애등록을 하지 않은 비율을 살펴보면 초졸과 고졸의 비율은 각각 2.3%와 2.4%로 유사하며 가장 높은 편이었고대졸 이상의 경우는 0.8%로 가장 낮았다. 최종 학력에 따른 장애등록 여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-20〉 최종 학력에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 최종 학력	예		아니오		전체
	빈도	행%	빈도	행%	
초졸	1,601	97.7%	37	2.3%	1,638
중졸	1,025	98.7%	14	1.3%	1,039
고졸	1,617	97.6%	39	2.4%	1,656

장애등록 여부 / 최종 학력	예		아니오		전체
	빈도	행%	빈도	행%	
대졸 이상	705	99.2%	6	0.8%	711
전체	4,948	98.1%	96	1.9%	5,044
참고: χ^2 = 8.913, P-value=0.030					

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-21〉은 장애인의 장애유형에 따른 장애등록여부 현황이다. 장애등록을 하지 않은 비율을 살펴보면 안면장애가 7.1%로 높은 편이었고 자폐성장애와 간장애의 경우 장애등록을 하지 않은 인원은 없는 것으로 나타났다. 장애 유형에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-21〉 장애유형에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 장애유형	예		아니오		전체
	빈도	행%	빈도	행%	
지체장애	3,157	98.7%	42	1.3%	3,199
뇌병변장애	623	96.9%	20	3.1%	643
시각장애	605	97.1%	18	2.9%	623
청각장애	800	95.6%	37	4.4%	837
언어장애	53	98.1%	1	1.9%	54
지적장애	463	98.1%	9	1.9%	472
자폐성장애	52	100%	0	0%	52
정신장애	181	95.3%	9	4.7%	190
신장장애	194	96.5%	7	3.5%	201
심장장애	26	96.3%	1	3.7%	27

장애유형 \ 장애등록 여부	예		아니오		전체
	빈도	행%	빈도	행%	
호흡기장애	51	96.2%	2	3.8%	53
간장애	55	100%	0	0%	55
안면장애	13	92.9%	1	7.1%	14
장루요루장애	81	96.4%	3	3.6%	84
뇌전증	43	95.6%	2	4.4%	45
전체	6,397	97.7%	152	2.3%	6,549
참고: χ^2 = 45.940, P-value=0.000					

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-22〉는 장애인 거주지의 시도에 따른 장애등록여부 현황을 보여주고 있다. 장애등록을 하지 않은 비율을 살펴보면 전체의 2.3%가 장애등록을 하지 않았다고 나타났다. 그중 부산이 7%로 높은 편에 속하였으나 울산과 세종의 경우 장애등록을 하지 않은 인원은 없는 것으로 나타났다. 시도에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하다고 할 수 있지만 뚜렷한 관계를 찾아보기는 힘들었다.

〈표 2-22〉 17개 시도별에 따른 장애등록여부

(단위: 명)

시도 \ 장애등록 여부	예		아니오		전체
	빈도	행%	빈도	행%	
서울	620	97.3%	17	2.7%	637
부산	384	93.0%	29	7.0%	413
대구	213	97.3%	6	2.7%	219
인천	290	98.6%	4	1.4%	294
광주	120	98.4%	2	1.6%	122

장애등록 여부 / 시도	예		아니오		전체
	빈도	행%	빈도	행%	
대전	232	98.7%	3	1.3%	235
울산	53	100%	0	0%	53
세종	41	100%	0	0%	41
경기	890	99.2%	7	0.8%	897
강원	324	97.0%	10	3.0%	334
충북	470	99.6%	2	0.4%	472
충남	347	98.9%	4	1.1%	351
전북	303	98.7%	4	1.3%	307
전남	806	97.1%	24	2.9%	830
경북	571	97.8%	13	2.2%	584
경남	588	95.8%	26	4.2%	614
제주	145	99.3%	1	0.7%	146
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 79.548, P-value=0.000

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

다음은 장애인의 거주지를 3개 권역으로 구분한 다음, 3개 권역에 따른 장애등록여부를 분석한 결과이다(〈표 2-23〉 참조). 장애등록을 하지 않은 비율을 살펴보면 서울, 광역, 세종이 3%로 높은 편인 반면에 중소도시는 1.4%로 낮은 편에 속하였다. 3개 권역에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-23〉 3개 권역에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 3개 권역	예		아니오		전체
	빈도	행%	빈도	행%	
서울, 광역, 세종	1,953	97.0%	61	3.0%	2,014
중소도시	2,526	98.6%	37	1.4%	2,563
농어촌	1,918	97.3%	54	2.7%	1,972
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 14.668, P-value=0.001

자료: 비공개 자료에 따른 출처 생략.

〈표 2-24〉는 장애인의 거주지를 3개 권역 및 동읍면으로 나누어 3개 권역 및 동읍면에 따른 장애등록여부 현황이다. 장애등록을 하지 않은 비율을 살펴보면 서울, 광역, 세종 및 동부가 3.1%로 높은 편에 속하였다. 그러나 서울, 광역, 세종 및 읍면부의 경우 장애등록을 하지 않은 인원은 없는 것으로 나타났다. 3개 권역 및 동읍면에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-24〉 3개 권역 및 동읍면에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 3개 권역 및 동읍면	예		아니오		전체
	빈도	행%	빈도	행%	
서울, 광역, 세종 및 동부	1,925	96.9%	61	3.1%	1,986
서울, 광역, 세종 및 읍면부	28	100%	0	0%	28
중소도시 및 동부	1,846	98.8%	23	1.2%	1,869
중소도시 및 읍면부	680	98.0%	14	2.0%	694
농어촌 및 읍면부	1,918	97.3%	54	2.7%	1,972
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 17.199, P-value=0.002

자료: 비공개 자료에 따른 출처 생략.

다음은 장애인 본인을 포함한 총 가구원 수에 따른 장애등록여부 현황에 대한 결과이다(〈표 2-25〉 참조). 장애등록을 하지 않은 비율을 살펴보면 총 가구원 수 1명인 경우 3.0%로 높은 반면에 4명인 경우(1.1%) 낮은 편에 속하였다. 본인을 포함한 총 가구원 수에 따른 장애등록여부의 차이는 유의수준을 5%로 가정한 경우 유의하지 않으나 그 차이가 매우 근소하게 나타났다.

〈표 2-25〉 본인을 포함한 총 가구원 수에 따른 장애등록여부

(단위: 명)

장애등록 여부 / 가구원 수	예		아니오		전체
	빈도	행%	빈도	행%	
1명	1,358	97.0%	42	3.0%	1,400
2명	2,674	97.5%	69	2.5%	2,743
3명	1,104	98.1%	21	1.9%	1,125
4명	777	98.9%	9	1.1%	786
5명 이상	484	97.8%	11	2.2%	495
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 9.145, P-value=0.058

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-26〉은 장애인 본인을 포함한 총 장애인 수에 따른 장애등록여부에 대한 현황이다. 장애등록을 하지 않은 비율을 살펴보면 총 장애인 수가 2명인 경우 3.6%로 높은 편에 속하였다. 1명, 3명 이상의 경우 2.1%, 2%로 유사하며 낮은 편이었고 본인을 포함한 총 장애인 수에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-26〉 본인을 포함한 총 장애인 수에 따른 장애등록여부

(단위: 명)

장애인 수 \ 장애등록 여부	예		아니오		전체
	빈도	행%	빈도	행%	
1명	5,366	97.9%	115	2.1%	5,481
2명	935	96.4%	35	3.6%	970
3명 이상	96	98.0%	2	2.0%	98
전체	6,397	97.7%	152	2.3%	6,549
참고: χ^2 = 8.324, P-value=0.016					

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 2-27〉은 장애인의 소득 유무에 따른 장애등록여부에 대한 현황이다. 장애등록을 하지 않은 비율을 살펴보면 소득이 없는 경우가 3%로 소득이 있는 경우(1%)보다 더 높은 편에 속하였다. 소득 유무에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-27〉 소득 유무에 따른 장애등록여부

(단위: 명)

소득 유무 \ 장애등록 여부	예		아니오		전체
	빈도	행%	빈도	행%	
소득 있음	2,208	99.0%	22	1.0%	2,230
소득 없음	4,189	97.0%	130	3.0%	4,319
전체	6,397	97.7%	152	2.3%	6,549
참고: χ^2 = 26.559, P-value=0.000					

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

다음은 장애인 가구의 월평균 총 가구소득을 4분위로 구분한 후 총 가구소득에 따른 장애등록여부에 대한 현황이다(〈표 2-28〉 참조). 장애등록을 하지 않은 비율을 살펴보면 1분위가 3.2%로 높은 편이고, 4분위가 1.8%로 낮은 편이었다. 월평균 총 가구소득에 따른 장애등록여부의 차이는 유의수준 5% 하에서 통계적으로 유의하게 나타났다.

〈표 2-28〉 월평균 총 가구소득에 따른 장애등록여부

(단위: 명)

장애등록여부 / 가구소득	예		아니오		전체
	빈도	행%	빈도	행%	
1분위	1,587	96.8%	53	3.2%	1,640
2분위	1,606	97.9%	35	2.1%	1,641
3분위	1,610	97.9%	35	2.1%	1,645
4분위	1,594	98.2%	29	1.8%	1,623
전체	6,397	97.7%	152	2.3%	6,549

참고: χ^2 = 8.570, P-value=0.036

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

가구원 특성 분석을 통해 유의한 결과를 가지는 설명변수를 사용한 포화모형에서 유의하지 않은 설명변수를 하나씩 제거하여 적합하였다. 〈표 2-29〉는 로지스틱 회귀분석의 최종 모형에 대한 결과이다. 최종 모형은 성별, 나이, 3개 권역 및 동읍면, 최종학력, 장애유형, 본인을 포함한 총 장애인 수의 6개 변수를 설명변수로, 종속변수는 장애등록여부로 하여 로지스틱 회귀모형을 적합하였다. 각 변수별 계수값, 표준오차 및 유의확률은 다음과 같다.

〈표 2-29〉 로지스틱 회귀분석 결과

설명변수	최종 모형		
	계수값	표준오차	유의확률
성별	.432	.215	.045
나이	.016	.009	.076
서울, 광역, 세종 및 동부			.085
서울, 광역, 세종 및 읍면부	-.823	.291	.005
중소도시 및 동부	-17.113	8395.316	.998
중소도시 및 읍면부	-.254	.359	.480
농어촌 및 읍면부	-.331	.259	.202
초졸			.065
중졸	-.336	.324	.299
고졸	.280	.264	.288
대졸 이상	-.701	.470	.135
지체장애			.023
뇌병변장애	1.038	.323	.001
시각장애	.602	.376	.109
청각장애	.986	.314	.002
언어장애	-16.390	6680.865	.998
지적장애	.243	.604	.687
자폐성장애	-15.924	6869.168	.998
정신장애	1.430	.435	.001
신장장애	1.169	.462	.011
심장장애	-16.721	8161.928	.998
호흡기장애	1.359	.756	.072
간장애	-16.585	5394.467	.998
안면장애	2.184	1.084	.044
장루요루장애	1.047	.748	.162
뇌전증	1.587	.771	.040
본인을 포함한 총 장애인수	.417	.203	.040
상수항	-5.800	.745	.000

자료: 비공개 자료에 따른 출처 생략.

지금까지 장애인실태조사에 참여한 가구원 6,549명에 대한 주요 특성 변수들 간의 연관성을 살펴보았다. 성별, 연령대, 최종 학력, 장애유형, 가구의 거주지를 17개 시도, 3개 권역, 3개 권역 및 동읍면으로 구분한 설명변수, 본인을 포함한 총 가구원 수, 본인을 포함한 총 장애인 수, 소득유무, 월평균 총 가구소득이 장애등록여부에 따라 통계적으로 유의하게 나타났다.

서울, 광역, 세종 및 동부 또는 농어촌 및 읍면부에 거주하며 성별이 여성이고 연령대는 10대 또는 70대 이상, 최종 학력은 초졸 또는 고졸, 안면장애를 가지고 있으며, 본인을 포함한 총 가구원 수가 1명 또는 2명이고 본인을 포함한 총 장애인 수가 2명이며 월평균 소득이 없고, 월평균 총 가구소득이 1분위인 경우 장애등록을 하지 않았을 비율이 높게 나타났다.

서울, 광역, 세종 및 읍면부에 거주하며 성별이 남성이고 연령대는 20대, 최종 학력은 대졸 이상, 자폐성장애 또는 간장애를 가지고 있으며, 본인을 포함한 총 가구원 수가 4명이고 본인을 포함한 총 장애인 수가 3명 이상이며, 월평균 소득이 있고 월평균 총 가구소득이 4분위인 경우 장애등록을 하지 않았을 비율이 낮게 나타났음을 확인할 수 있었다.

제4절 소결

이 장에서는 장애인실태조사의 표본설계 현황을 파악하였고, 장애인실태조사의 목적 중 하나인 장애인 출현율의 정의 및 현황에 대해서도 살펴보았다. 또한, 2017년 실태조사 분석을 통하여 장애인 유무 가구 및 장애 등록 여부에 따른 주요 특성 변수들 간 연관성을 도출하였다.

제1절은 1995년부터 2017년까지의 조사 연도별 표본설계 현황에 대해 살펴보았다. 목표모집단은 장애인이고, 표본추출틀은 2008년(등록장애인 DB 사용)을 제외한 나머지 조사 연도에서는 인구센서스(인구주택총조사) 기반 조사구를 사용하였다. 인구센서스 기반 조사구에는 장애유형 관련 정보가 없어서 표본설계 시 이를 반영할 수 없는 상황이다. 대신 표본배분 단계에서 제한적 정보이지만 지역별 장애인 분포를 사용하여 보완하고 있다. 그래서 시도별로 장애인 출현율이 낮은 특정 장애인을 포괄할 수 있도록 하고 있다. 표본추출은 이중추출방법으로, 1상에서는 가구 및 판별조사를 대규모로 실시하고, 1상에서의 판별여부에 따라 2상에서는 심층조사를 실시하고 있다.

제2절의 내용을 요약하면 2017년 전체 장애인(재가 및 시설장애인을 모두 포함)은 2,668,411명으로 1995년(1,053,468명) 이후로 계속 증가하는 추세이나, 전체 장애인 출현율은 2017년에는 5.39%로 2011년(5.61%)이후 조금씩 감소하고 있음을 확인할 수 있다. 또한 1995년의 경우 2.35%로 가장 낮은 장애인 출현율을 가진다. 전체 장애인 중에서 재가장애인이 시설장애인에 비해 높은 비중을 차지하고 있어서, 재가장애인 출현율과 전체 장애인 출현율 간 차이가 크지 않다. 한편, 2008년은 등록 장애인만을 대상으로 조사했기 때문에 장애인 출현율은 산출되지 않았다.

조사연도별 장애유형별 장애인 출현율 변화 추이는 모든 조사연도별에서 지체장애 출현율이 가장 높으며 다음으로 뇌병변장애, 시각장애, 청각장애 출현율이 높은 편에 속한다. 한편 중복장애 출현율은 2005년, 2000년, 1995년에는 지체장애 출현율 다음으로 높은 것으로 나타났으며 나머지 장애유형 출현율은 높지 않은 편이었다.

제3절은 2017년 장애인실태조사 자료를 가지고 가구에 대한 주요 특성 변수들 간 연관성을 살펴본 결과 원조사구이면서 농어촌 및 읍면부의 단독주택에 거주하며 본인을 포함한 총 가구원 수가 6명 이상이고, 월평균 총 가구소득이 1분위인 가구에 장애인이 있을 비율이 높게 나타났다. 또한 가구원에 대한 분석 결과는 서울, 광역, 세종 및 동부 또는 농어촌 및 읍면부에 거주하며 성별이 여성이고 연령대는 10대 또는 70대 이상이며 최종 학력은 초졸 또는 고졸로 안면장애를 가지고 있으며, 본인을 포함한 총 가구원 수가 1명 또는 2명이고 본인을 포함한 총 장애인 수가 2명이며 월평균 소득이 없고, 월평균 총 가구소득이 1분위인 경우 장애등록을 하지 않았을 비율이 높게 나타났다.

제3장

표본설계 효율화를 위한 다각적 접근

제 3 장 표본설계 효율화를 위한 다각적 접근

제1절 표본조사구의 축소

가장 최근에 실시한 2017년 장애인실태조사 자료를 사용하여 표본조사구 축소에 대한 모의실험을 실시하였다. 먼저 표본조사구의 현황을 살펴보았다.

1. 2017년 표본조사구 현황

2017년 조사의 경우 모집단 층화는 2016년 등록장애인 수 분포를 고려하여 권역별(3개 층, 대도시/중소도시/농어촌)로 적정 표본 규모를 산정하였다. 그 다음 전국 17개 시도별로 층화하였고, 7개 특별·광역시를 제외한 9개 도 및 세종특별자치시 지역은 동부와 읍면부로 구분하여 3차 층화를 하였다. 그런 다음 조사구 특성에 따라 표본조사구를 추출하였다. 즉 층화는 권역 → 시도 → 읍면으로 구분된다고 볼 수 있다.

〈표 3-1〉에서 보듯이 중소도시(39.8%)와 대도시(38%)가 표본조사구의 대부분을 차지하고 있으며 농어촌은 22.2%로 나타났다.

〈표 3-1〉 권역별 표본조사구 분포

권역	표본조사구	
	개	%
대도시	760	38.0
중소도시	797	39.8
농어촌	444	22.2
전체	2,001	100.0

자료: 비공개 자료에 따른 출처 생략.

〈표 3-2〉는 2017년 표본조사구 분포로 총 2,001개의 표본조사구를 사용하였다. 〈표 2-3〉의 2017년에 산출한 1,000개와 다른 까닭은 2015년 등록센서스 기반 표본추출틀 조사구 중에서 병합조사구를 활용해야 하는데, 그 당시 제공시점이 10월로 예정되어 있어서 개별조사구를 2개 병합하여 평균 60가구로 만들어 활용했기 때문이다. 동부는 서울특별시와 경기도가 19.9%로 표본조사구 비중이 가장 높았으며 다음으로 부산광역시(9.3%), 인천광역시(8.2%), 대구광역시(5.8%), 대전광역시(5.8%) 순으로 나타났다. 이에 반해 울산광역시(1.2%)와 세종특별자치시(0.6%)는 동부 중에서 아주 낮은 편에 속하였다. 읍면부는 전라남도가 20.2%, 경상북도는 18.9%, 경상남도가 14%로 높은 편이고 세종특별자치시(1.3%), 전라북도(2.5%), 제주도(2.6%)가 낮은 편으로 나타났다.

〈표 3-2〉 시도별 표본조사구 분포

(단위: %)

지역	동부		읍면부		전체	
	개	%	개	%	개	%
서울특별시	272	19.9	-	-	272	13.6
부산광역시	128	9.3	-	-	128	6.4
대구광역시	80	5.8	-	-	80	4.0

지역	동부		읍면부		전체	
	개	%	개	%	개	%
인천광역시	112	8.2	-	-	112	5.6
광주광역시	56	4.1	-	-	56	2.8
대전광역시	80	5.8	-	-	80	4.0
울산광역시	16	1.2	-	-	16	0.8
세종특별자치시	8	0.6	8	1.3	16	0.8
경기도	272	19.9	72	11.4	344	17.2
강원도	40	2.9	57	8.9	97	4.8
충청북도	40	2.9	56	8.8	96	4.8
충청남도	32	2.4	72	11.4	104	5.2
전라북도	56	4.1	16	2.5	72	3.6
전라남도	48	3.5	128	20.2	176	8.8
경상북도	48	3.5	120	18.9	168	8.4
경상남도	56	4.1	88	14.0	144	7.2
제주도	24	1.8	16	2.6	40	2.0
전국	1,368	100.0	633	100.0	2,001	100.0

자료: 비공개 자료에 따른 출처 생략.

다음은 장애인 가구를 포함하고 있는 표본조사구의 분포에 대해 전체 장애인 가구, 미등록 장애인 가구로 나누어 살펴보았다. 여기서 전체 장애인은 등록 장애인과 미등록 장애인을 모두 포함하는 것이다. 표본조사구에서 할당된 가구를 접촉하기 전에는 가구 내 법정장애인의 포함여부를 알 수 없다. 장애인실태조사는 법정장애인이 있는 가구의 경우 가구 및 판별조사뿐만 아니라 심층조사도 작성하기 때문에 표본조사구 내에 법정장애인을 포함하고 있는 가구 현황을 살펴보는 것은 의미가 있다. 더불어 미등록 장애인 가구가 포함되어 있는 표본조사구 분포도 파악하였다.

〈표 3-3〉을 보면 전체 표본조사구에서 전체 장애인 가구를 포함하고

있는 표본조사구 비중은 전국의 경우 91.5%로 나타나 대부분의 표본조사구에서 전체 장애인 가구를 포함하고 있음을 알 수 있다. 전국 동읍면부별로 보면 읍면부가 95.3%로 동부(89.7%)에 비해 5.6%p 높은 편에 속한다. 시도별 및 동부·읍면부별로 보면 세종특별자치시의 동부가 62.5%로 낮은 편에 속하나, 나머지 지역은 85% 이상으로 나타나 표본조사구의 대부분은 전체 장애인 가구를 포함하는 것으로 파악된다.

〈표 3-3〉 전체 장애인 가구를 포함하고 있는 표본조사구에 대한 비중

(단위: %)

지역	비중		
	동부	읍면부	전체
서울특별시	86.4	-	86.4
부산광역시	89.8	-	89.8
대구광역시	96.3	-	96.3
인천광역시	89.3	-	89.3
광주광역시	87.5	-	87.5
대전광역시	90.0	-	90.0
울산광역시	100.0	-	100.0
세종특별자치시	62.5	87.5	75.0
경기도	87.5	94.4	89.0
강원도	90.0	100.0	95.9
충청북도	97.5	92.9	94.8
충청남도	87.5	94.4	92.3
전라북도	92.9	93.8	93.1
전라남도	95.8	97.7	97.2
경상북도	87.5	92.5	91.1
경상남도	100.0	95.5	97.2
제주도	87.5	100.0	92.5
전국	89.7	95.3	91.5

주: 비중=(전체 장애인 가구를 포함하고 있는 표본조사구 수/전체 표본조사구 수)×100
자료: 비공개 자료에 따른 출처 생략.

〈표 3-4〉는 전체 표본조사구에서 미등록 장애인 가구를 포함하고 있는 표본조사구의 비중을 보여주고 있다. 전국 전체로 보면 6.5%로 낮은 편이고, 동읍면별로 보면 읍면부가 8.2%로 동부(5.6%) 보다 2.6%p 높게 나타났다. 시도별 및 동부·읍면부별로 보면 부산광역시의 동부가 16.4%, 경상남도의 읍면부가 15.9%, 전라남도의 읍면부가 12.5%로 미등록 장애인 가구를 포함하고 있는 조사구가 높은 편에 속하였다. 이에 반해 울산광역시 동부, 세종특별자치시 동부 및 읍면부, 충청남도 동부, 전라북도 읍면부, 제주도 동부는 미등록 장애인을 포함하는 표본조사구가 없는 것으로 나타났다.

〈표 3-4〉 미등록 장애인 가구를 포함하고 있는 표본조사구에 대한 비중

(단위: %)

지역	비중		
	동부	읍면부	전체
서울특별시	6.6	-	6.6
부산광역시	16.4	-	16.4
대구광역시	6.3	-	6.3
인천광역시	3.6	-	3.6
광주광역시	3.6	-	3.6
대전광역시	3.8	-	3.8
울산광역시	0.0	-	0.0
세종특별자치시	0.0	0.0	0.0
경기도	2.6	2.8	2.6
강원도	12.5	8.8	10.3
충청북도	2.5	1.8	2.1
충청남도	0.0	5.6	3.9
전라북도	7.1	0.0	5.6
전라남도	6.3	12.5	10.8

지역	비중		
	동부	읍면부	전체
경상북도	2.1	7.5	6.0
경상남도	5.4	15.9	11.8
제주도	0.0	6.3	2.5
전국	5.6	8.2	6.5

주: 비중=(미등록 장애인 가구를 포함하고 있는 표본조사구 수/전체 표본조사구 수)×100
자료: 비공개 자료에 따른 출처 생략.

다음은 표본조사구 내 전체 및 미등록 장애인 규모로 〈표 3-5〉와 같다. 전체 장애인의 경우 경기도(14%), 전라남도(12.4%), 서울특별시(10.1%)가 높은 편이었으며, 세종특별자치시(0.6%) 및 울산광역시(0.8%)가 낮은 편으로 나타났다. 미등록 장애인의 경우 부산광역시(18.5%), 경상남도(16.6%), 전라남도(15.3%)가 높은 편으로 나타나 전체 장애인과는 다른 결과를 보였다.

〈표 3-5〉 표본조사구 내 전체 및 미등록 장애인 규모

(단위: %)

지역	전체 장애인			미등록 장애인		
	동부	읍면부	전체	동부	읍면부	전체
서울특별시	17.0	-	10.1	21.3	-	12.1
부산광역시	10.3	-	6.1	32.6	-	18.5
대구광역시	5.6	-	3.3	6.7	-	3.8
인천광역시	7.7	-	4.5	4.5	-	2.5
광주광역시	3.2	-	1.9	2.2	-	1.3
대전광역시	6.2	-	3.7	3.4	-	1.9
울산광역시	1.4	-	0.8	0.0	-	0.0
세종특별자치시	0.3	1.0	0.6	0.0	0.0	0.0
경기도	16.9	9.6	14.0	7.9	2.9	5.7

지역	전체 장애인			미등록 장애인		
	동부	읍면부	전체	동부	읍면부	전체
강원도	2.8	8.3	5.1	5.6	8.8	7.0
충청북도	4.8	10.4	7.1	1.1	1.5	1.3
충청남도	2.5	9.9	5.5	0.0	5.9	2.5
전라북도	6.0	2.6	4.6	4.5	0.0	2.5
전라남도	4.9	23.4	12.4	3.4	30.9	15.3
경상북도	3.4	16.9	8.9	1.1	17.6	8.3
경상남도	5.2	15.0	9.1	5.6	30.9	16.6
제주도	1.8	2.9	2.3	0.0	1.5	0.6
전국	100.0	100.0	100.0	100.0	100.0	100.0

자료: 비공개 자료에 따른 출처 생략.

앞의 결과는 2017년 실태조사의 표본조사구 현황이기에 일반화할 수는 없으나 현황 파악 시 참고 자료로 활용하는 데 의미가 있다.

2. 표본조사구 축소 관련 모의실험

장애인 출현율을 위한 가구조사의 비중을 줄일 수 있는지를 검토하기 위하여 모의실험을 실시하였다. 가구조사는 표본설계를 통해 추출된 표본조사구 내의 가구와 접촉하여 조사를 하는 방식이다. 모의실험은 2017년 조사의 2,001개 표본조사구를 모집단으로 설정하였다. 표본조사구의 축소 비율을 5가지(10%, 20%, 30%, 40%, 50%)로 설정하여 다양한 축소 비율에 따른 표본조사구의 축소 효과를 비교하였다.

모의실험 절차는 표본조사구를 10% 축소하는 경우를 가지고 설명하면 다음과 같다. 모집단 표본조사구를 전국 17개 시도별로 층화한 후 7개의 특별·광역시(서울특별시, 부산광역시, 대구광역시, 인천광역시, 광주광역

시, 대전광역시, 울산광역시)를 제외한 9개의 도(경기도, 강원도, 충청북도, 충청남도, 전라북도, 전라남도, 경상북도, 경상남도, 제주도)와 세종특별자치시는 동부와 읍면부로 층화하였다. 각 해당하는 셀(cell) 마다 10%를 축소한 값을 구하였고, 이 값이 추출해야 하는 표본조사구의 개수가 되는 것이다. 예를 들면 서울의 경우 모집단의 표본조사구는 272개(〈표 3-2〉참조)인데 10%를 축소한다면 추출해야 할 표본조사구는 245개(272×0.9=244.8)이고, 세종특별자치시의 경우도 동일한 방법으로 계산하면 7개(8×0.9=7.2)인 것이다. 10%를 축소하는 경우 표본조사구의 추출 개수는 총 1,798개(동부: 1,230개, 읍면부: 568개)이다(〈표 3-6〉참조).

시도별 동부·읍면부별로 해당하는 표본조사구 개수를 단순임의추출(simple random sampling) 방법으로 최종 추출하였는데 모의실험의 반복수는 100번으로 설정하였다.

20%를 축소하는 경우 표본조사구의 추출 개수는 총 1,602개(동부: 1,095개, 읍면부: 507개)이다([부록 1] 참조). 30%를 축소하는 경우 표본조사구의 추출 개수는 총 1400개(동부: 957개, 읍면부: 443개)이다([부록 2] 참조). 40%를 축소하는 경우 표본조사구의 추출 개수는 총 1,203개(동부: 822개, 읍면부: 381개)이다([부록 3] 참조). 50%를 축소하는 경우 표본조사구의 추출 개수는 총 1,001개(동부: 684개, 읍면부: 317개)이다([부록 4] 참조).

〈표 3-6〉 표본조사구를 10% 축소할 경우 지역별 동부·읍면부별 표본조사구 추출 개수

(단위: 개)

지역	동부	읍면부	전체
서울특별시	245	-	245
부산광역시	115	-	115
대구광역시	72	-	72
인천광역시	101	-	101
광주광역시	50	-	50
대전광역시	72	-	72
울산광역시	14	-	14
세종특별자치시	7	7	14
경기도	245	65	310
강원도	36	51	87
충청북도	36	50	86
충청남도	29	65	94
전라북도	50	14	64
전라남도	43	115	158
경상북도	43	108	151
경상남도	50	79	129
제주도	22	14	36
전체	1,230	568	1,798

자료: 비공개 자료에 따른 출처 생략.

모의실험 결과의 평가지표는 단순임의추출 가정으로 95% 신뢰수준 하에서 가구수 기준 최대허용오차 한계이며 식은 다음과 같다. 최대허용오차는 표본오차[1](sampling error)에 대해 허용 가능한 최댓값을 의미하며, 표본오차는 표준오차(standard error)와 상대표준오차(Relative Standard Error: RSE) 등으로 보여준다.

1) 표본오차란 모집단 전수조사가 아닌 표본조사를 실시하기 때문에 발생하는 오차임.

$$\text{최대허용오차} = 1.96 \times \frac{0.5}{\sqrt{\text{해당셀의가구수}}} \times 100$$

최대허용오차 한계는 모의실험이 1번 실행될 때마다 생성한 값을 누적한 후 평균값으로 제시하였다. 오차의 한계가 클수록 구간이 넓어지며 점추정치에 대해 확신할 수 없다고 볼 수 있다. 권역별, 시도별 동부·읍면부별 최대허용오차 한계를 구하였고, 이와 함께 가구조사 대상 가구 분포, 전체 장애인 및 미등록 장애인 분포도 함께 살펴보았다.

〈표 3-7〉을 보면 표본조사구 축소 30% 이하의 경우 권역별 가구수 기준 최대허용오차는 모두 1% 내외의 값을 가진다. 표본조사구 축소 40% 이상에서는 모두 1% 이상으로 나타났다. 표본조사의 공표 범위는 권역별 기준(대도시, 중소도시, 농어촌)으로 하고 있다.

다음은 시도별 동부·읍면부별 가구수 기준 최대허용오차 한계에 대해 살펴보았다(〈표 3-8〉 참조). 전국의 경우 표본조사구 축소 비율과 상관없이 모두 1% 이하로 나타났다. 표본조사구 축소 30% 이하에서는 시도별 동부·읍면부별 최대허용오차가 모두 10% 이하의 값을 가졌다. 그러나 표본조사구 축소 40% 이상의 경우 세종특별자치시 동부 및 읍면부에서 10% 이상으로 나타났다. 그 외 나머지 지역에서는 모두 10% 이하로 나타났다.

다음은 조사구 축소 비율에 따른 시도별 동부·읍면부별 전체 장애인 규모(〈표 3-9〉 참조)와 모집단 대비 미등록 장애인 비중([그림 3-1] 참조)을 살펴보았다. 전체 장애인 규모를 살펴보면 모집단은 6,820명이다. 표본조사구를 10%로 축소한 경우, 전체 장애인 규모는 6,121명으로 나타났으며 모집단 대비 89.8%를 차지하였다. 즉, 표본조사구를 10%로 축소하면 조사대상자 규모도 약 10% 축소된다고 볼 수 있다. 표본조사구를 20%로 축소한 경우에는 5,460명(80.1%), 30%의 경우 4,777명(70.0%),

40%의 경우 4,098명(60.1%)이고 50%의 경우 3,416명(50.1%)으로 나타났다(〈표 3-9〉 참조). 모집단 대비 미등록 장애인 비중은 다음과 같다. 표본조사구를 10%로 축소한 경우에는 모집단 대비 미등록 장애인의 비중은 89.2%로 나타났다. 이는 표본조사구를 10% 축소하면 미등록 장애인의 규모도 10.2% 축소된다고 볼 수 있다. 표본조사구를 20% 축소한 경우에는 모집단 대비 미등록 장애인의 비중은 80.3%, 30%의 경우 70.1%, 40%의 경우 59.9%이고 50%의 경우 51.0%로 나타났다([그림 3-1] 참조).

이상의 모의실험 결과를 보듯이 표본조사구를 축소한 비율만큼 조사대상 규모도 비슷한 비율로 축소되는 것을 확인하였다. 추후 조사방법을 변경한다면 조사의 정확도 및 비용을 종합적으로 고려하여 최대허용오차 기준을 결정하기 위한 심층연구를 실시해야 할 것이다.

〈표 3-7〉 권역별 가구수 기준 최대허용오차 한계

(단위: 가구)

권역	모집단		표본조사구 축소 10%		표본조사구 축소 20%		표본조사구 축소 30%		표본조사구 축소 40%		표본조사구 축소 50%	
	가구수	최대 허용 오차	가구수	최대 허용 오차	가구수	최대 허용 오차	가구수	최대 허용 오차	가구수	최대 허용 오차	가구수	최대 허용 오차
대도시	13,745	0.836	12,348	0.882	11,001	0.934	9,618	0.999	8,266	1.078	6,870	1.182
중소도시	14,455	0.815	13,002	0.859	11,584	0.911	10,119	0.974	8,701	1.051	7,209	1.154
농어촌	8,000	1.096	7,170	1.157	6,402	1.225	5,591	1.311	4,801	1.415	4,016	1.547

자료: 비공개 자료에 따른 출처 생략.

〈표 3-8〉 시도별 동부·읍면부별 가구수 기준 최대허용오차 한계

(단위: 가구)

지역		모집단		표본조사구 축소 10%		표본조사구 축소 20%		표본조사구 축소 30%		표본조사구 축소 40%		표본조사구 축소 50%	
		가구수	최대허용오차	가구수	최대허용오차	가구수	최대허용오차	가구수	최대허용오차	가구수	최대허용오차	가구수	최대허용오차
전국		36,200	0.515	32,520	0.543	28,987	0.576	25,328	0.616	21,767	0.664	18,095	0.729
서울특별시	동부	4,918	1.397	4,430	1.472	3,946	1.560	3,438	1.671	2,948	1.805	2,459	1.976
부산광역시	동부	2,316	2.036	2,080	2.149	1,846	2.281	1,626	2.431	1,394	2.625	1,158	2.881
대구광역시	동부	1,447	2.576	1,300	2.718	1,158	2.880	1,012	3.081	867	3.329	724	3.644
인천광역시	동부	2,026	2.177	1,826	2.294	1,628	2.429	1,411	2.609	1,214	2.813	1,012	3.082
광주광역시	동부	1,013	3.079	904	3.260	814	3.434	705	3.692	614	3.954	506	4.356
대전광역시	동부	1,447	2.576	1,303	2.715	1,158	2.880	1,012	3.080	868	3.326	723	3.646
울산광역시	동부	290	5.755	254	6.153	236	6.383	198	6.962	181	7.294	145	8.141
세종특별자치시	동부	144	8.167	126	8.736	108	9.420	108	9.420	91	10.296	72	11.556
	읍면부	144	8.167	126	8.736	107	9.488	107	9.488	89	10.408	72	11.632
경기도	동부	4,941	1.394	4,451	1.469	3,961	1.557	3,452	1.668	2,962	1.801	2,471	1.972
	읍면부	1,300	2.718	1,173	2.862	1,048	3.027	904	3.260	775	3.522	648	3.851
강원도	동부	724	3.642	652	3.839	579	4.071	509	4.346	436	4.695	361	5.162
	읍면부	1,021	3.067	912	3.245	823	3.416	717	3.661	608	3.975	519	4.303

지역		모집단		표본조사구 축소 10%		표본조사구 축소 20%		표본조사구 축소 30%		표본조사구 축소 40%		표본조사구 축소 50%	
		가구수	최대허용 오차	가구수	최대허용 오차	가구수	최대허용 오차	가구수	최대허용 오차	가구수	최대허용 오차	가구수	최대허용 오차
충청북도	동부	722	3.647	650	3.844	578	4.077	505	4.361	433	4.708	360	5.166
	읍면부	1,007	3.088	900	3.267	809	3.447	703	3.698	612	3.964	502	4.378
충청남도	동부	584	4.055	529	4.259	475	4.499	403	4.883	346	5.268	292	5.734
	읍면부	1,298	2.720	1,172	2.863	1,046	3.030	902	3.264	776	3.519	648	3.849
전라북도	동부	1,009	3.085	901	3.265	810	3.443	702	3.698	612	3.961	505	4.360
	읍면부	288	5.775	252	6.177	233	6.416	198	6.967	181	7.298	144	8.181
전라남도	동부	867	3.328	777	3.516	687	3.740	615	3.953	524	4.282	432	4.714
	읍면부	2,311	2.039	2,074	2.152	1,841	2.284	1,626	2.430	1,390	2.629	1,155	2.884
경상북도	동부	871	3.321	780	3.508	689	3.734	616	3.950	526	4.272	437	4.690
	읍면부	2,165	2.106	1,949	2.220	1,732	2.355	1,513	2.519	1,301	2.717	1,080	2.982
경상남도	동부	1,020	3.068	911	3.247	821	3.421	710	3.679	621	3.932	509	4.343
	읍면부	1,597	2.452	1,432	2.589	1,271	2.749	1,126	2.921	961	3.161	797	3.471
제주도	동부	436	4.693	399	4.905	344	5.282	307	5.594	254	6.150	216	6.672
	읍면부	294	5.715	258	6.101	240	6.333	203	6.883	183	7.241	148	8.068

자료: 비공개 자료에 따른 출처 생략.

〈표 3-9〉 시도별 동부·읍면부별 전체 장애인 규모

(단위: 명)

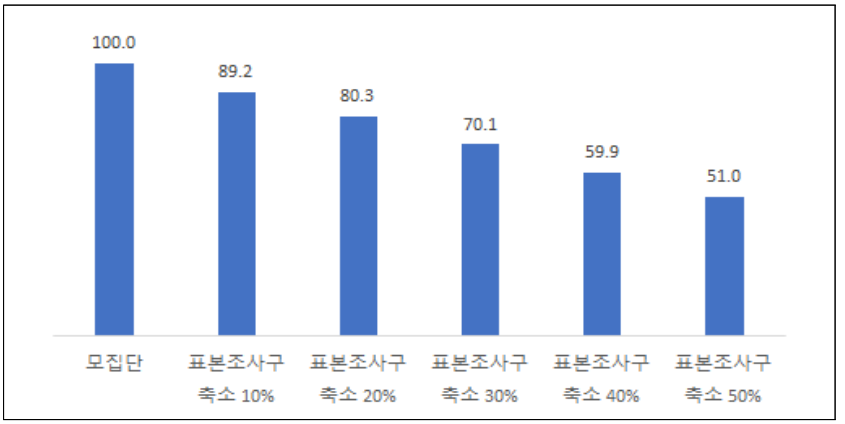
지역		모집단	표본조사구 축소 10%	표본조사구 축소 20%	표본조사구 축소 30%	표본조사구 축소 40%	표본조사구 축소 50%
전국		6,820	6,121	5,460	4,777	4,098	3,416
서울특별시	동부	688	619	552	480	411	345
부산광역시	동부	418	374	333	294	252	209
대구광역시	동부	228	205	183	158	138	114
인천광역시	동부	310	279	248	216	186	156
광주광역시	동부	129	114	104	90	79	64
대전광역시	동부	250	225	199	176	150	125
울산광역시	동부	55	48	45	38	34	28
세종특별자치시	동부	14	12	11	11	9	7
	읍면부	28	25	21	21	17	15
경기도	동부	686	618	550	479	411	341
	읍면부	266	240	215	185	158	134
강원도	동부	115	104	93	81	69	58
	읍면부	230	206	185	161	136	117
충청북도	동부	195	176	157	139	117	98
	읍면부	289	258	232	202	175	144
충청남도	동부	102	93	83	71	61	51
	읍면부	274	247	220	190	164	137
전라북도	동부	243	217	193	166	146	123
	읍면부	71	63	58	49	44	36
전라남도	동부	199	178	158	141	119	99
	읍면부	647	580	514	456	387	324
경상북도	동부	137	123	109	97	83	69
	읍면부	468	421	374	327	282	235
경상남도	동부	210	187	169	147	128	105
	읍면부	414	372	330	294	251	206
제주도	동부	73	67	57	51	43	37
	읍면부	81	71	66	57	51	40

주: 모집단의 전국 장애인이 6,820명인 이유는 가구 및 판별조사를 통해 파악된 규모이고, 이 중에서 심층조사에 6,594명이 응답한 것임.

자료: 비공개 자료에 따른 출처 생략.

〔그림 3-1〕 조사구 축소 비율별 모집단 대비 미등록 장애인의 비중 - 전국

(단위: %)



자료: 비공개 자료에 따른 출처 생략.

제2절 이중추출틀을 활용한 표본추출 방법론

1. 이중추출틀 활용에 대한 이론적 고찰[2)]

가. 이론적 고찰

하나의 모집단에 대해 표본추출틀(sampling frame)이 두 개 이상 존재할 수 있다. 예를 들어 개인단위 조사에서 개인의 휴대전화 리스트와 개인이 속한 가구의 가구 전화번호 리스트가 그것이다. 한편 이러한 추출틀들은 비교적 완전한 추출틀로 있지만, 불완전한 리스트도 많이 존재한다. 불완전한 추출틀의 원인을 살펴보면 실제로 존재하는 단위가 추출틀

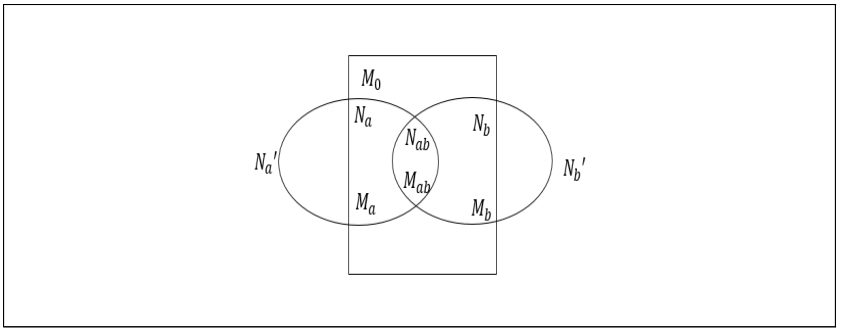
2) 본문의 내용은 박홍래(2006)와 Arcos, Molina, Ranalli, and Rueda (2015)의 내용을 참고하여 편집함.

명부에서 누락되거나, 실제로 존재하지 않는 단위가 리스트에 허위로 기재되어 있는 경우 등으로 인한 것이다. 이러한 사례로서는 사업체 조사에서 폐업 사업체가 추출틀 명부에 존재하거나, 신규 사업체가 표본추출틀 명부에 기재되지 않고 누락되는 경우가 해당된다. 또한 동일한 단위가 중복되어 2회 이상 기재되어 있는 경우도 추출틀이 불완전하게 되는 원인이라고 할 수 있다.

동일 모집단에 대해 2개의 추출틀을 사용하여 표본추출할 경우 표본의 정도를 높일 수 있으며, 비용도 절약할 수 있다. 하지만 2개 추출틀을 사용하게 되면 조사단위를 실제보다 많이 포함시킬 위험이 있기 때문에 주의해야 한다.

두 개의 추출틀 A, B 구조를 [그림 3-2]와 같이 표현할 수 있다.

[그림 3-2] 이중추출틀 구조



자료: Arcos, A., Molina, D., Ranalli, M. G., & del Mar Rueda, M. (2015). Frames2: A Package for Estimation in Dual Frame Surveys. p.53 내용을 바탕으로 수정

N은 추출틀에 포함된 단위의 총수이며, M은 모집단의 총 단위수라 하자. 그러면 추출틀 A의 총수는 $N_A = N_a + N_{ab} + N_a'$이다. 이때 N_a'는 추출틀 A에 포함되어 있으나, 실제로는 존재하지 않는 단위의 수를 나타낸다. N_{ab}는 추출틀 A와 B에 동시에 포함된 단위의 수이며, N_a는 추출틀 A에만 존재

하는 단위의 수를 나타낸다. 추출틀 B에 대해서도 $N_B = N_b + N_{ab} + N_b'$가 되며, 추출틀 A에서 정의된 단위들의 수와 유사하게 정의할 수 있다.

한편 모집단의 총 단위 수는 $M = M_a + M_{ab} + M_b + M_0$가 된다. M_0는 모집단에는 포함되어 있으나, 두 개의 추출틀 A와 B에서 누락된 단위들의 수를 나타낸다. M개의 단위가 모두 리스트에 포함된 경우 즉, $M_0 = 0$인 경우를 고려하여 추정식을 유도하고자 한다. 또한 $N_a' = 0$이고, $N_b' = 0$인 경우를 가정한다. 즉, $N_A = M_A$, $N_B = M_B$인 경우이다.

n_a와 n_b는 각 영역으로부터 추출된 표본크기이며, μ_a, μ_b, $\bar{y}_a$, $\bar{y}_b$는 각각 두 영역에 대한 모평균과 표본평균을 의미한다. 또한 n_{ab}'과 n_{ab}''은 추출틀 A와 추출틀 B에서 추출된 A, B 공통 영역의 표본크기를 나타낸다. 이때 N_a, N_b, N_{ab}는 알고 있다고 가정한다.

그러면 모집단의 관심변수를 y라 하면, y_k는 k번째 단위의 관찰값이다 ($k = 1,2,\cdots,N$). 유한 모집단 총합이 $\tau_y = \Sigma_k y_k$이며, 이를 각 영역의 총합으로 표현하면 다음과 같다.

$$Y = Y_a + Y_{ab} + Y_b$$

여기서 $Y_a = \sum_{k \in a} y_k$, $Y_{ab} = \sum_{k \in ab} y_k$, $Y_b = \sum_{k \in b} y_k$이다.

추출틀 A와 B로부터 각각 크기가 n_A, n_B인 표본 s_A와 s_B를 추출한다고 하면, 추출틀 A에 있는 단위 k가 표본에 포함될 확률은 $\pi_k^A = \Pr(k \in s_A)$, 추출틀 B에 있는 단위 k가 표본에 포함될 확률은 $\pi_k^B = \Pr(k \in s_B)$가 된다.

표본 s_A로부터 관찰된 자료를 이용하여 추출틀 A에서 각 영역에 대한 총합 추정량은 각각 다음과 같다.

$$\hat{Y}_a = \sum_{k \in s_A} \delta_k(a) d_k^A y_k, \quad \hat{Y}_{ab}^A = \sum_{k \in s_A} \delta_k(ab) d_k^A y_k$$

여기서 만일 $k \in a$이면 $\delta_k(a) = 1$, 그렇지 않으면 0이며, 또한 $k \in ab$이면 $\delta_k(ab) = 1$, 그렇지 않으면 0이다. 한편 d_k^A는 추출틀 A를 이용한 추출설계에 대한 가중치이며, 포함확률의 역수로서 $d_k^A = 1/\pi_k^A$이다.

마찬가지로 표본 s_B로부터 관찰된 자료를 이용하여 추출틀 B에서 각 영역에 대한 총합 추정량은 각각 다음과 같다.

$$\hat{Y}_b = \sum_{k \in s_B} \delta_k(b) d_k^B y_k, \quad \hat{Y}_{ab}^B = \sum_{k \in s_B} \delta_k(ab) d_k^B y_k$$

여기서 만일 $k \in b$이면 $\delta_k(b) = 1$, 그렇지 않으면 0이며, 또한 $k \in ab$이면 $\delta_k(ab) = 1$, 그렇지 않으면 0이다. 한편 d_k^B는 추출틀 B를 이용한 추출설계에 대한 가중치이며, 포함확률의 역수로서 $d_k^B = 1/\pi_k^B$이다.

이중추출틀을 이용한 조사에서 모집단 총합에 대한 추정방법은 여러 연구자별로 다르게 제시하고 있는데, Hartley(1962)는 $\hat{Y}_{ab}^A$와 $\hat{Y}_{ab}^B$에 대해 가중치 θ를 사용하여 다음과 같이 제시하였다.

$$\hat{Y}_H = \hat{Y}_a + \theta \hat{Y}_{ab}^A + (1-\theta) \hat{Y}_{ab}^B + \hat{Y}_b$$

여기서 $0 \le \theta \le 1$이다.

한편 Hartely(1974)는 가중치 θ의 최적값을 다음과 같이 구하였다.

$$\theta_{opt} = \frac{V(\hat{Y}_{ab}^A) + Cov(\hat{Y}_b, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}$$

즉, θ의 최적값은 결과적으로 표본설계에 대해 추정량의 분산을 최소로 한다.

이 값은 모집단 분산과 공분산을 이용하기 때문에 실제로는 계산이 불가능하며, 따라서 표본자료를 이용하여 추정할 수 있다.

한편 $\hat{Y}_H$의 분산추정치는 추출틀 A와 B로부터 각각 독립적으로 표본을 추출한 것으로 고려하면 다음과 같이 구할 수 있다.

$$\hat{V}(\hat{Y}_H) = \hat{V}(\hat{Y}_a) + \theta^2 \hat{V}(\hat{Y}_{ab}^A) + \theta \widehat{Cov}(\hat{Y}_a, \hat{Y}_{ab}^A) + (1-\theta)^2 \hat{V}(\hat{Y}_{ab}^B) + \hat{V}(\hat{Y}_b) + (1-\theta) \widehat{Cov}(\hat{Y}_b, \hat{Y}_{ab}^B)$$

Fuller and Burmeister(1972)는 두 추출틀 A와 B의 중복된 영역에 대한 크기를 추정하여 다음과 같은 총합 추정량을 제안하였다.

$$\hat{Y}_{FB} = \hat{Y}_a + \hat{Y}_b + \beta_1 \hat{Y}_{ab}^A + (1-\beta_1) \hat{Y}_{ab}^B + \beta_2 (\hat{N}_{ab}^A - \hat{N}_{ab}^B)$$

여기서 $\hat{N}_{ab}^A = \sum_{k \in s_A} \delta_k(ab) d_k^A$, $\hat{N}_{ab}^B = \sum_{k \in s_B} \delta_k(ab) d_k^B$이며, β_1과 β_2는 다음과 같이 추정한다.

$$\begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} = - \begin{bmatrix} V(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B) & Cov(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B) \\ Cov(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B) & V(\hat{N}_{ab}^A - \hat{N}_{ab}^B) \end{bmatrix}^{-1} \times \begin{bmatrix} Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}_{ab}^B, \hat{Y}_{ab}^A - \hat{Y}_{ab}^B) \\ Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}_{ab}^B, \hat{N}_{ab}^A) \end{bmatrix}$$

이때 추정량의 분산을 최소로 하는 최적값은 β_1과 β_2 를 의미한다.

표본을 통해 추정된 모수 $\hat{\beta}_1$와 $\hat{\beta}_2$를 이용하여 총합 추정량 $\hat{Y}_{FB}$의 분산 추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_{FB}) = \hat{V}(\hat{Y}_a) + \hat{V}(\hat{Y}_B) + \hat{\beta}_1(\widehat{Cov}(\hat{Y}_a, \hat{Y}_{ab}^A) - \widehat{Cov}(\hat{Y}_B, \hat{Y}_{ab}^B)) + \hat{\beta}_2(\widehat{Cov}(\hat{Y}_a, \hat{N}_{ab}^A) - \widehat{Cov}(\hat{Y}_B, \hat{N}_{ab}^B))$$

여기서 $\hat{Y}_B = \hat{Y}_b + \hat{Y}_{ab}^B$이다.

이중추출틀을 이용한 추정에서 Bankier(1986)와 Kalton and Anderson(1986)은 두 개의 추출틀로부터 추출되는 표본 s_A와 s_B로 모든 추출단위를 한 개의 추출틀로부터 추출된 하나의 표본으로 결합하는 방법을 제안하였다. 이 경우 중복된 단위들에 대해서는 추출가중치가 필요하며 편향제거를 위해 다음과 같이 가중치를 조정한다.

$$\tilde{d}_k^A = \begin{cases} d_k^A, & \text{if } k \in a \\ (1/d_k^A + 1/d_k^B)^{-1}, & \text{if } k \in ab \end{cases},$$

$$\tilde{d}_k^B = \begin{cases} d_k^B, & \text{if } k \in B \\ (1/d_k^A + 1/d_k^B)^{-1}, & \text{if } k \in ab \end{cases}$$

또는 이를 정리하면, 다음과 같이 표현할 수 있다.

$$\tilde{d}_k = \begin{cases} d_k^A, & \text{if } k \in A \\ (1/d_k^A + 1/d_k^B)^{-1}, & \text{if } k \in ab \\ d_k^B, & \text{if } k \in B \end{cases}$$

이와 같이 조정된 가중치를 적용한 총합추정치는 다음과 같다.

$$\hat{Y}_{BKA} = \sum_{k \in s_A} \tilde{d}_k^A y_k + \sum_{k \in s_B} \tilde{d}_k^B y_k = \sum_{k \in s} \tilde{d}_k y_k$$

여기서 $s = s_A \cup s_B$이다.

이 추정량을 계산하기 위해서는 표본추출설계 하에서의 추출틀이 중복되는 영역에 포함될 확률과 두 추출틀에 단위가 속하는 포함확률을 모두 알아야 한다.

만일 추출틀의 크기 N_A와 N_B를 알 수 있다면, 레이킹비(raking ratio) 조정과 같은 방법으로 효율성을 높일 수 있다. 레이킹비 조정을 통해 조정된 추정량은 다음과 같다.

$$\hat{Y}_{SFR} = \frac{N_A - \hat{N}^r_{ab}}{\hat{N}_a}\hat{Y}^A_a + \frac{N_{B-}\hat{N}^r_{ab}}{\hat{N}_b\hat{Y}^B_b} + \frac{\hat{N}^r_{ab}}{\hat{N}_{abS}}\hat{Y}_{abS}$$

여기서 $\hat{Y}_{abS} = \sum_{k\in s_A}\tilde{d}^A_k\delta_k(ab)y_k + \sum_{k\in s_B}\tilde{d}^B_k\delta_k(ab)y_k$,

$\hat{N}_{abS} = \sum_{k\in s_A}\tilde{d}^A_k\delta_k(ab) + \sum_{k\in s_B}\tilde{d}^B_k\delta_k(ab)$

$\hat{N}_a = \sum_{k\in s_A}\delta_k(a)$, $\hat{N}_b = \sum_{k\in s_B}\delta_k(b)$ 이며 $\hat{N}^r_{ab}$은 2차방정식

$\hat{N}_{abS}x^2 - (\hat{N}_{abS}(N_A+N_B) + \hat{N}^A_{aS}\hat{N}^B_{bS})x + \hat{N}_{abS}N_AN_B = 0$의 해이다.

한편 Skinner and Rao(1996)은 Fuller and Burmeister(1972)가 제안한 최대우도추정량(maximum likelihood estimator)을 복합표본설계에 확장한 유사최대우도추정량(pseudo maximum likelihood estimator)을 다음과 같이 제안하였다.

$$\hat{Y}_{PML} = \frac{N_A - \hat{N}^{PML}_{ab}(\gamma)}{\hat{N}^A_a}\hat{Y}^A_a + \frac{N_B - \hat{N}^{PML}_{ab}(\gamma)}{\hat{N}^B_b}\hat{Y}^B_b + \frac{\hat{N}^{PML}_{ab}(\gamma)}{\gamma\hat{N}^A_{ab} + (1-\gamma)\hat{N}^B_{ab}}\left[\gamma\hat{Y}^A_{ab} + (1-\gamma)\hat{Y}^B_{ab}\right]$$

여기서 $\hat{N}^{PML}_{ab}(\gamma)$는 2차방정식 최소의 해이며, $\gamma\in(0,1)$이다.

한편 상수 γ는 다음과 같이 구할 수 있으며, 이는 $\hat{Y}_{PML}$의 분산을 최소로 한다.

$$\gamma_{opt} = \frac{\hat{N}_a N_B V(\hat{N}_{ab}^B)}{\hat{N}_a N_B V(\hat{N}_{ab}^B) + \hat{N}_b N_A V(\hat{N}_{ab}^A)}$$

그러면 $\hat{Y}_{PML}$의 분산추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_{PML}) = \hat{V}(\sum_{k \in s_A} \tilde{z}_k^A) + \hat{V}(\sum_{k \in s_B} \tilde{z}_k^B)$$

여기서 만일 $k \in a$이면 $\tilde{z}_k^A = y_k - \frac{\hat{Y}_a}{\hat{N}_a}$이며, 만일 $k \in ab$이면

$\tilde{z}_k^A = \hat{\gamma}_{opt}\left(y_k - \frac{\hat{Y}_{ab}^A}{\hat{N}_{ab}^A}\right) + \hat{\lambda}\hat{\phi}$이다.

또한 $\hat{\gamma}_{opt} = \frac{n_A/N_A \hat{Y}_{ab}^A + n_B/N_B \hat{Y}_{ab}^B}{n_A/N_A \hat{N}_{ab}^A + n_B/N_B \hat{N}_{ab}^B} - \frac{\hat{Y}_a}{\hat{N}_a} - \frac{\hat{Y}_b}{\hat{N}_b}$이며, $\hat{\phi} = \frac{n_A \hat{N}_b}{n_A \hat{N}_b + n_B \hat{N}_a}$이다.

이와 유사하게 만일 $k \in a$이면 $\tilde{z}_k^B = y_k - \frac{\hat{Y}_b}{\hat{N}_b}$이며, 만일 $k \in ab$이면

$\tilde{z}_k^B = (1-\hat{\gamma}_{opt})\left(y_k - \frac{\hat{Y}_{ab}^B}{\hat{N}_{ab}^B}\right) + \hat{\lambda}(1-\hat{\phi})$이다.

최근에 Ranalli, Arcos, Rueda and Teodoro(2013)은 일종의 보조정보를 이용할 수 있다는 가정 하에서 이중추출틀을 이용한 추정에 보정추정(calibration estimation)방법을 적용하였다.

p차원의 보조변수벡터를 $\boldsymbol{x}_k = (x_{1k}, x_{2k}, \cdots, x_{pk})$로 가정하면 단위 k에 대한 보조변수값이 존재한다. 각각의 보조변수는 추출틀 A, 추출틀 B, 또는 전체 모집단에 있는 단위들에 대해 이용가능하다. 이와 더불어 보조변수

의 모집단 총합 $X=\sum_{k\in U}\boldsymbol{x}_k$는 알고 있다고 가정한다. 그러면 이중표본 추출틀에 대한 보정추정량은 다음과 같이 정의할 수 있다.

$$\hat{Y}_{CAL1}=\sum_{k\in s}d_k^{CAL1}y_k$$

여기서 d_k^{CAL1}은 $\sum_{k\in s}d_k^{CAL1}\boldsymbol{x}_k=X$의 조건 하에서 거리함수 $\sum_{k\in s}G(d_k^{CAL1},\check{d}_k)$를 최소로 하는 새로운 보정가중치이며, 기존의 가중치 $\check{d}_k$는 다음과 같다.

$$\check{d}_k=\begin{cases}d_k^A, & \text{if } k\in a\\ \eta d_k^A, & \text{if } k\in ab\cap s_A\\ (1-\eta)d_k^B, & \text{if } k\in ab\cap s_B\\ d_k^B, & \text{if } k\in b\end{cases}$$

이때, $\eta\in[0,1]$이다.

그러면 또 다른 $\hat{Y}_{BKA}$에 대한 보정추정량은 다음과 같이 정의할 수 있다.

$$\hat{Y}_{CAL2}=\sum_{k\in s}d_k^{CAL2}y_k$$

여기서 가중치 d_k^{CAL2}는 $\sum_{k\in s}d_k^{CAL2}\boldsymbol{x}_k=X$의 조건 하에서 거리함수 $\sum_{k\in s}G(d_k^{CAL2},\tilde{d}_k)$를 최소로 하는 새로운 보정가중치이며, 기존의 가중치 $\tilde{d}_k$는 다음과 같다.

$$\tilde{d}_k=\begin{cases}d_k^A, & \text{if } k\in A\\ (1/d_k^A+1/d_k^B)^{-1}, & \text{if } k\in ab\\ d_k^B, & \text{if } k\in B\end{cases}$$

임의의 보정추정량에 대한 분산추정량은 Deville(1993)의 방법을 이용하여 다음과 같이 구할 수 있다.

$$\hat{V}(\hat{Y}) = \frac{1}{1-\Sigma_{k\in s}a_k^2}\sum_{k\in s}\frac{d_k^*-1}{d_k^*}\left(d_k^* e_k - \sum_{j\in s} a_j d_j^* e_j\right)^2$$

여기서 d_k^*는 $\hat{Y}_{CAL1}$ 또는 $\hat{Y}_{CAL2}$ 정의한 $\check{d}_k$ 또는 $\tilde{d}_k$가 될 수 있다. 한편 $a_k = \frac{d_k^*-1}{d_k^*} / \sum_{j\in s}\frac{d_j^*-1}{d_j^*}$이며, e_k는 일반화회귀에서의 잔차를 나타낸다.

특정한 추정량에 의존한 분산추정방법은 추정량 간 비교를 통해 잘못된 결과를 도출할 가능성이 있다. 이를 보완하기 위해 추정량과 무관하게 동일한 형태의 분산추정방법 중의 하나가 Quenoullie(1949, 1956)에 의해 제안된 잭나이프 분산추정량이다.

층화를 고려하지 않은 각 추출틀로부터 수집된 자료에 대해 임의의 추정량 $\hat{Y}_c$에 대한 잭나이프 분산추정량은 다음과 같다.

$$\hat{V}_J(\hat{Y}) = \frac{n_A-1}{n_A}\sum_{i\in s_A}(\hat{Y}_c^A(i)-\overline{Y}_c^A)^2 + \frac{n_B-1}{n_B}\sum_{j\in s_B}(\hat{Y}_c^B(j)-\overline{Y}_c^B)^2$$

여기서 $\hat{Y}_c^A(i)$는 표본 s_A로부터 단위 i를 제거한 후 계산된 추정량 $\hat{Y}_c$값이며, $\overline{Y}_c^A$는 $\hat{Y}_c^A(i)$의 평균이며, 이와 유사하게 $\hat{Y}_c^B(j)$는 표본 s_B로부터 단위 j를 제거한 후 계산된 추정량 $\hat{Y}_c$ 값이며, $\overline{Y}_c^B$는 $\hat{Y}_c^B(j)$의 평균이다.

만일 각 표본 추출틀이 층으로 구성되어 있다고 가정할 경우를 고려해 보자. 즉, 추출틀 A가 H의 층으로 구성되고, 추출틀 B 또한 L개의 층으로 구성되어 있다고 가정하자. 그러면 추출틀 A로부터 h층에 대해 모집단 단위 N_{Ah}에서 n_{Ah}개의 표본 단위를 추출하게 된다, 마찬가지로 추출틀 B에 대해 l층에서 모집단 단위 N_{Bl}에서 n_{Bl}개의 표본 단위를 추출한다면, 잭나이프 분산추정량은 다음과 같다.

$$\hat{V}_J(\hat{Y}) = \sum_{h=1}^{H} \frac{n_{Ah}-1}{n_{Ah}} \sum_{i \in s_{Ah}} (\hat{Y}_c^A(hi) - \overline{Y}_c^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl}-1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{Y}_c^B(lj) - \overline{Y}_c^{Bl})^2$$

여기서 $\hat{Y}_c^A(hi)$는 표본 s_{Ah}로부터 단위 i를 제거한 후 계산된 추정량 $\hat{Y}_c$값이며, $\overline{Y}_c^{Ah}$는 $\hat{Y}_c^A(hi)$의 평균이며, 이와 유사하게 $\hat{Y}_c^B(lj)$는 표본 s_{Bl}로부터 단위 j를 제거한 후 계산된 추정량 $\hat{Y}_c$ 값이며, $\overline{Y}_c^{Bl}$는 $\hat{Y}_c^B(lj)$의 평균이다.

한편 보다 일반적인 상황으로 층화집락(stratified cluster) 설계를 고려할 수 있다. 추출틀 A는 H개의 층으로 구성되고, h층은 N_{Ah}개의 관찰단위로 구성되며, $\tilde{N}_{Ah}$는 1차 추출단위(PSU: primary sampling units)인 집락으로 구성되어 있으며, $\tilde{n}_{Ah}$개의 집락을 추출하게 된다.

마찬가지로 추출틀 B는 L개의 층으로 구성되고, l층은 N_{Bl}개의 관찰단위로 구성되며, $\tilde{N}_{Bl}$은 1차 추출단위인 집락으로 구성되어 있으며, $\tilde{n}_{Bl}$개의 집락을 추출하게 된다.

잭나이프 분산추정량을 정의하기 위해 $\tilde{Y}_c^A(hj)$를 추출틀 A에서 h층의 j번째 PSU를 제거하고 구한 추정량 $\hat{Y}_c$이라 하고, 이와 유사하게 $\tilde{Y}_c^B(lk)$ 또한 추출틀 B에서 l층의 k번째 PSU를 제거하고 구한 추정량 $\hat{Y}_c$이라 하자. 그러면 잭나이프 분산추정량은 다음과 같이 정의할 수 있다.

$$\hat{V}_J(\hat{Y}_c) = \sum_{h=1}^{H} \frac{\tilde{n}_{Ah}-1}{\tilde{n}_{Ah}} \sum_{j=1}^{\tilde{n}_{Ah}} (\tilde{Y}_c^A(hj) - \tilde{Y}_c^{Ah})^2 + \sum_{l=1}^{L} \frac{\tilde{n}_{Bl}-1}{\tilde{n}_{Bl}} \sum_{k=1}^{\tilde{n}_{Bl}} (\tilde{Y}_c^B(lk) - \tilde{Y}_c^{Bl})^2$$

여기서 $\tilde{Y}_c^{Ah}$는 $\tilde{Y}_c^{Ah}(hj)$들의 평균이며, $\tilde{Y}_c^{Bl}$는 $\tilde{Y}_c^{Bl}(lk)$들의 평균이다.

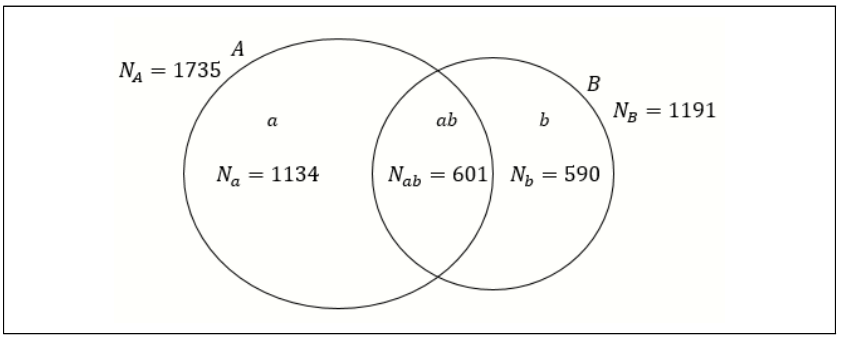
나. 데이터 분석 사례

이중추출틀에 대한 데이터 분석 사례를 설명하기 위해 먼저 R 패키지의 Frame2 패키지를 이용하도록 한다. 이 패키지는 이중 표본추출틀에 대한 점추정과 구간추정을 하는 새로운 R패키지이다.

이 패키지에는 8개의 주요 함수(Hartley, FB, SFRR, PML, PEL, CalSF, CalDF)를 포함하며, 이 함수들은 모두 추정량을 구하는 함수들이다. 한편 이 패키지에는 "Compare"라는 추가적인 함수를 포함하며, 이는 모든 가능한 추정량들에 대한 요약 통계를 보여준다. 또한 6개의 부가적인 함수를 이용하여 보조적인 결과를 유도할 수 있는데, HT(Horvitz-Thompson) 추정량의 계산, 또는 두 HT 추정량 간의 공분산 등의 계산을 하는 데 활용할 수 있다. 한편으로 이 패키지에는 8개의 함수와 더불어 잭나이프 분산 추정량의 계산과 신뢰구간을 제공한다.

함수가 어떻게 수행되는지를 알아보기 위해 패키지에 있는 DatA, DatB의 두 데이터셋을 이용한다. DatA에는 추출틀 A로부터 $N_A = 1{,}735$개의 가구에서 $n_A = 105$개의 가구를 층화추출설계로 추출된 표본에 관한 정보가 있다. 추출틀 A는 6개의 층으로 구분되며, 각 층은 $N_{Ah} = (727, 375, 113, 186, 115, 219)$로부터 단순임의 표본 $n_{Ah} = (15, 20, 15, 20, 15, 20)$을 추출한다. 다른 한편으로 추출틀 B는 $N_b = 1{,}191$개 가구로 구성되며, 이중에서 단순임의 표본 $n_B = 135$가구를 추출한다. 또한 중복된 영역에 대해 $N_{ab} = 601$개 가구로 구성된다고 하자([그림 3-3] 참조).

〔그림 3-3〕 이중추출틀에 대한 데이터 분석 사례



자료: Arcos, A., Molina, D., Ranalli, M. G., & del Mar Rueda, M. (2015). Frames2: A Package for Estimation in Dual Frame Surveys. p.60 내용을 바탕으로 수정

보조정보가 있는 경우, 또는 보조정보가 없는 경우 등에 대해 총합 추정을 위해 다양한 함수를 이용할 수 있다.

또 다른 사례로서 'Dat'에는 실제 전화조사에 대한 이중추출틀을 고려할 수 있는데 스페인의 안달루시안 주민들에게 이민에 관한 여론조사를 실시한 데이터를 포함한다. 이 조사에서는 두 개의 표본추출틀을 이용하고 있는데, 하나는 유선전화(landline phone)번호이고 다른 하나는 휴대전화(cell phone)번호이다. 유선전화 추출틀에서는 크기가 1,919개인 층화 표본을 추출하였고, 휴대전화번호에서는 483개의 단순임의 표본을 이용하였다. 1차 포함확률은 유선전화 추출틀에서 층화임의 추출설계로부터 계산하였고, 유선전화수와 성인수를 이용하여 조정하였다. 휴대전화 추출틀에 대해서 1차 포함확률은 개인별로 주어진 휴대전화수로 계산하고 수정하였다. 자료수집 시점에서 유선전화와 휴대전화의 추출틀 크기는 각각 4,982,920개와 5,707,655개였으며 전체 모집단 크기는 6,350,916개였다.

이러한 설계를 이용하여 R패키지에서 각 추출틀의 포함확률을 계산하고, 추정량을 산출하고, 분산 추정량을 계산하게 된다.

2. 희귀집단 판별조사를 위해 RDD 조사에서의 이중추출틀 설계[3)]

가. 배경

정부 기관을 대상으로 한 몇 가지 대규모 조사는 무작위 전화걸기(RDD: Random Digit Dialing) 표본을 이용해 전화로 실시한다. 예를 들어 행동 위험 요소 감시 시스템, 국가면역조사(NIS: National Immunization Survey), 주 및 지방통합 전화조사가 해당된다. 이러한 조사의 일부는 특정 대상 모집단에 초점을 맞춘다. 왜냐하면 주요 대상개체들이 모집단의 특정한 부분과 관련이 있기 때문이다.

예로는 65세 이상 노인, 0~17세 아동, 천식 환자, 특별한 건강관리가 필요한 사람 등이 해당될 수 있다. 이러한 잠재적 목표 모집단중 일부의 경우 적격 가구의 비율은 상당히 낮을 수 있으며 관련 조사에서는 대규모 RDD 가구 표본을 통해 적격 가구여부를 판별하여 대상자인지를 구분하게 된다. 대규모 RDD 샘플의 전화번호는 상당한 노력을 필요로 한다. 가구의 적격여부에 대한 판단을 위해 대규모 조사원들이 필요하며, 조사부터 판별까지 많은 시간이 소요될 수 있다. 조사 수행에 드는 상당한 비용은 대상개체가 얼마나 드문지 또는 최종 표본에 적격가구 표본이 얼마나 필요한지에 따라 결정된다.

희귀집단을 표본 추출할 때 데이터 수집 비용을 줄이기 위해 여러 가지 표본 설계가 있다. Kalton and Anderson(1986)은 희귀집단의 특성을 추정하기 위해 희귀집단이 집중된 층수를 과대표집할 것을 제안하였다. Srinath(2002)는 판별 표본 크기를 최소화하는 층에 대한 할당과 과대표집에 의한 정밀도의 손실을 고려했다. 다른 설계 옵션에는 주거용 전화번

3) 본문의 내용은 Srinath, Battaglia, and Khare(2004)의 연구를 참고하여 편집함.

호 목록을 과대표집하여 주거용 전화번호의 높은 비율을 포함하는 표본을 표집했다. 그리고 목록보조 RDD 표본 설계에서 작동하는 전화번호 집단의 기준을 높여 표본에서 비작동 번호의 비율을 감소시키는 네트워크(현장적응) 표본추출 및 이중 표본추출틀이 옵션에 포함된다.

예를 들어, Brick, Judkins & Morganstein(2002)은 2상 설계를 사용하여 전화번호를 서로 다른 작동 중인 주거용 번호 비율로 층화하도록 제안한다.

이중추출틀 접근방식으로 RDD 조사에 더 많은 관심을 받기 시작하고 있다. 이중 표본추출틀의 가장 단순한 적용으로는 가구의 완전한 추출틀을 보유하고 있지만, 그 추출틀에서 적격가구여부에 대해서는 식별되지 않는 경우이다.

두 번째 추출틀은 대상 모집단에 속하는 가구만 포함하지만, 전체 대상 모집단을 포괄하지는 않는다. 이중추출틀을 이용한 표본추출의 기본 아이디어는 각각 두 추출틀에서 표본을 추출하는 것이다. 조사는 각 표본에 대해 실시되며 각각 가중치가 부여된다. 그런 다음 전체 추출틀과 부분 추출틀의 중복(겹침)과 관련된 정보를 사용하여 두 표본을 함께 사용하며 편향되지 않은 추정치를 제공할 수 있는 가중치를 개발한다. 이는 이중추출틀 추출의 중요한 측면으로, 일부 적격 가구는 전체 추출틀에만 존재하는 반면 다른 가구는 전체 추출틀과 부분 추출틀에 모두 존재하기 때문에 다중 선택 가능성이 있기 때문이다.

나. 국가면역조사에서의 이중추출틀

국가면역조사는 미국의 질병통제예방센터(CDC: Centers for Disease Control and Prevention)에 의해 실시되는 대규모 RDD 조사이다. 대상자는 만19~35개월 아동의 예방접종률을 측정한다. 하지만 미국의

19~35개월 자녀를 둔 가구의 비율은 전체가구의 약 4% 미만인 것으로 나타나고 있다. 이러한 사실은 하나의 적격가구로 식별하기 위해 약 25가구 이상의 표본가구를 접촉해야 한다는 것이다. 적격가구로의 판정단계와 면접단계에서의 무응답효과는 판정에 필요한 가구의 표본크기를 증가시키게 된다. 또한 RDD 표본에는 작동하지 않는 전화번호와 비거주(사업체) 전화번호와 같이 이외의 전화번호가 포함될 수 있다. 또한 해당 전화번호가 주거용인지, 사업용인지, 아니면 비업무용 전화번호인지 전혀 결정되지 않는 번호인 미결정 전화번호가 존재하게 되면 전화걸기를 시도해야 할 전화번호의 수는 더욱 증가하게 된다.

따라서 국가면역조사에서는 표본선정을 위해 미국전체를 주, 도시지역(도시 또는 카운티) 및 나머지 주지역(보스턴과 매사추세츠-잔여지역을 2개의 층으로) 등으로 총 78개 층으로 층화하였다.

이러한 층화는 면역 행동 계획(IAP) 지역으로 알려져 있다. 각 분기마다 목록보조(list-assisted) RDD 표본을 각 층에서 추출한다. 목록보조 표본설계에서는 주거용 주소 목록에 있는 전화번호가 0인 100개의 연속 전화번호의 묶음(bank)은 제외한다(제로뱅크(zero-bank)로 표시). 각 분기별 RDD 표본은 표본공지와 관리목적으로 부차표본(복제)으로 구분된다. 각 IAP 지역에는 분기당 목표 면접수가 있기 때문에 표본은 통제된 방식으로 공지된다. 이러한 목표 면접수는 IAP 지역당 평균 110개 가구에 대한 면접이다. 전화걸기과정의 첫 번째 단계는 그 전화번호가 알려진 가구인지 확인하는 것이다. 그런 다음 주거용 전화번호를 선별하여 해당 가구에 19~35개월의 자녀가 한 명 이상 있는지 여부를 판단한다. 만약 그 가구가 적격대상 가구이면, 자녀의 예방접종에 대해 가장 잘 알고 있는 가구원과의 면접이 시도된다. 위에서 언급한 바와 같이, 무응답은 이 과정의 세 단계에서 발생한다. 첫째, 어떤 전화번호는 절대 해결되지 않는다. 미해결 전화번호의 일부는 연령 적격자녀가 포함된 가구일 수도 있

다. 둘째, 알려진 가구 중 일부는 연령 적격성을 판단하기 위해 판별 면접을 완료하지 않을 수 있다. 이 가구들 중 일부는 적격 연령아동을 포함하고 있을 수 있다. 셋째, 확인된 적격가구 중 일부는 예방접종에 대한 면접을 완료하지 못할 수도 있다. 일부 연령대 대상 가구는 자녀가 없음을 표시(단순거절)해 스스로 걸러낼 가능성도 있다. 이 모든 요인은 NIS의 초기 RDD 표본 크기와 관련이 있으며 NIS에서 걸려온 RDD 판별 통화 수를 감소시키는 데 있어 매우 중요한 관심사가 된다.

2002년에는 30,974건의 가구면접이 완료되었다. 필요한 초기 RDD 표본크기는 3,361,396개의 전화번호였다. 면접원은 총 2,055,371개의 전화번호에 전화걸기를 시도하였고, 이중에서 1.5%의 적격률을 나타냈다. 적격률은 완전하게 면접한 전화번호수를 전화걸기한 총 전화번호수로 나눈 값이다.

2003년 동안 이중추출틀 설계의 잠재적 리스트에 대한 조사가 이루어졌다. 기본 아이디어는 RDD 표본 구성 요소를 유지하고, 리스트 표본을 설계에 추가하는 것이었다. RDD 추출틀은 비전화 가구와 제로 뱅크의 가구를 제외한 전체 목표 모집단을 포괄한다. 리스트 추출틀(전화 번호 포함)은 적격 전화 가구에 대한 부분적인 포괄성과 제로 뱅크의 적격 전화 가구 중 일부에 대한 포괄성도 제공할 것이다. 그러나 리스트 추출틀은 매우 불완전할 수 있다. 즉, 이러한 리스트 추출틀이 목표 모집단의 일부분만 포함할 수도 있다.

일부 리스트 추출틀은 더 나은 포괄성을 제공할 수 있지만, 리스트정보의 품질은 떨어질 수 있다. 리스트 추출틀에 포함된 전화번호를 사용하려고 할 때, 리스트의 최신화는 특별히 관심을 가져야 한다. 마케팅 시스템 그룹과 협력하여 어린 자녀가 있는 가구에 적용되는 다양한 잠재적 상업용 전화번호 리스트를 검토하였다. 어린 자녀가 있는 가구를 대상으로 하는 'Experian New Babies(ENB)' 리스트를 가장 유망한 목록으로 파

악했다. ENB 목록은 어린 자녀가 있는 가구에 마케팅을 하고자 하는 기업을 대상으로 한다. 이 목록에는 이름과 주소 정보를 포함하고 있는데, 사전에 우편을 부치는 데 사용할 수 있다. 이 목록은 주별로 출생 기록과 기타 출처로부터 작성된다. 목록의 일부에는 전화번호도 포함되어 있다. 그 목록은 다양한 출처의 정보를 사용하여 최신으로 유지된다. ENB 리스트와 같은 목록은 주마다 포괄성과 품질이 다를 수 있지만, 이것은 식별 가능한 이중추출틀 설계의 가장 좋은 리스트 출처로 고려하였다.

이러한 ENB 리스트의 적격률은 RDD의 적격율보다 훨씬 더 높을 것으로 기대하며 목록에 전화번호가 있는 가구의 초기 주 단위 수치에 따르면 대상 모집단의 포괄성이 30~40%에 달할 수 있을 것으로 예상되었다.

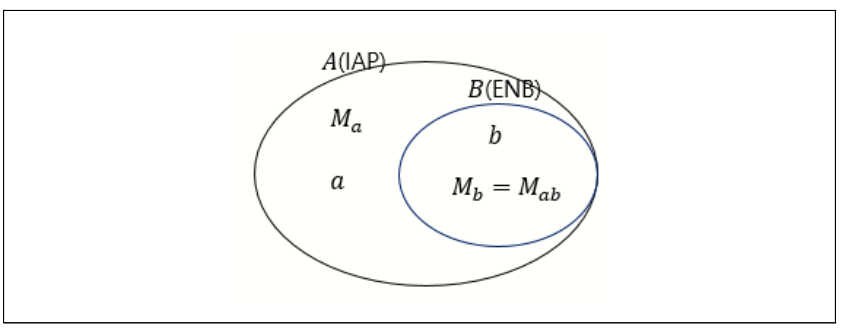
2003년 4분기에 이중추출틀에 대한 테스트가 시행되었다. RDD 샘플은 통상적인 방식으로 5개 도시 IAP 지역에 선정되었다. GENESYS Sampling Systems와 협력하여 ENB 목록에 있는 전화번호 목록에서 5개의 층에 있는 각각 전화번호의 간단한 무작위 표본을 선택했다. RDD 표본과 ENB표본의 전화번호 매칭을 시도하였고, 중복된 전화번호는 RDD 표본에 남겨두었다. 그 후 표본에 포함된 각각 중복 사례의 최종 처분을 ENB 표본으로 변환하였다. 이는 마치 두 개의 분리된 표본 전화번호로 전화하는 것을 피하기 위해서였다.

이중추출틀 설계를 위한 가중치를 개발하려면 RDD 표본(완료된 추출틀에서 나온 것으로 가정)과 ENB 목록 추출틀(대상 모집단의 일부를 포함하도록 가정) 사이의 중복도를 결정해야 한다. 이는 각 층에 대한 RDD 표본을 추출하여 해당 층에 대한 전체 ENB 추출틀(전화번호가 있는 연령 적격 가구)과 전화번호 표본을 일치시킴으로써 가능하다.

다. 이중추출틀에 대한 표본 할당

전화번호 목록으로 두 개의 표본추출틀 A와 B가 있다고 하자([그림 3-4] 참조). M_a개 전화번호는 추출틀 A에만 속하도록 하고, M_b개의 전화번호는 추출틀 B에만 속하며, M_{ab} 개의 전화번호는 추출틀 A와 B에 모두 속한다고 하자. 그러면 NIS는 RDD 조사이기 때문에 추출틀 A는 IAP 지역이나 주에서 선택할 수 있는 모든 전화번호로 구성되어 있어 100% 포괄성이 있으며, 따라서 $M_{ab} = M_b$ 이다. 추출틀 B는 추출틀 A의 부분집합이며, 만 19~35개월의 자녀를 둔 가구의 ENB 목록에 있는 전화번호로 구성된다. $M = M_a + M_{ab}$는 추출틀 A에서 선택할 수 있는 총 전화번호 수를 나타낸다. m_a와 $m_b = m_{ab}$이 알려져 있다. 미지의 N은 추출틀 A에서 19~35개월 사이의 자녀를 둔 가구의 전화번호 수를 나타낸다.

[그림 3-4] NIS조사에서의 이중추출틀



자료: Arcos, A., Molina, D., Ranalli, M. G., & del Mar Rueda, M. (2015). Frames2: A Package for Estimation in Dual Frame Surveys. p.53 내용을 바탕으로 수정

$e = N/M$은 본 추출틀에서 적격률이라 하자. 마찬가지로, $N_b = N_{ab}$은 추출틀 B에서 19개월에서 35개월 사이의 자녀를 둔 주거용 전화번호 수를 가리킨다.

$e_b = N_b/M_b$는 추출틀 B의 적격률을 나타내며, e_b는 ENB 추출틀의 오차로 인해 1보다 작을 것이다. 단, e_b가 e보다 훨씬 클 것으로 예상한다.

추출틀 A의 총 숫자 모집단 중 $\alpha = M_b/M$을 ENB 리스트가 포괄하는 전화번호의 비율이 되도록 하자. 그러면 $(1-\alpha) = M_a/M$ 이다.

N_a는 추출틀 A에는 있지만 ENB 리스트 상에는 없는 M_a개의 전화번호 중 19개월에서 35개월 사이의 자녀를 둔 가구의 수를 나타낸다. $e_a = N_a/M_a$는 ENB 리스트에 없는 추출틀의 일부에 대한 적격률이며, $e = (1-\alpha)e_a + \alpha e_b$가 된다.

추출틀 A에서 m개의 전화번호를 단순임의 추출하고, 추출틀 B에서도 m_b개의 전화번호를 단순임의 추출한다고 가정하자. 선택된 총 전화번호 표본은 $m_0 = m + m_b$ 개이다. n은 표본으로 추출된 m 개의 전화번호 중 19개월에서 35개월 사이의 자녀를 둔 가구의 수를 나타낸다. 한편 가구당 19개월에서 35개월 사이의 아이가 한 명 있다고 가정하면 n_b는 m_b개의 전화번호로 전화한 결과 추출틀 B에서 자녀가 있는 가구의 수를 나타낸다. 선택된 총 자녀수는 $n_0 = n + n_b$ 가 된다.

전체 ENB 목록에 포함된 전화번호는 n개의 적격 가구 전화번호 표본과 일치시킨다. n_{ab} 개의 전화번호는 두 표본에 나타난 일부 전화번호가 포함되어 있게 된다. 그러면 $n_a = n - n_{ab}$로써 ENB 목록상에서 일치하지 않는 번호를 갖는 가구들의 수이다. 따라서 여기에서는 3개의 표본을 정의할 수 있다. 크기 n의 표본은 추출틀 A에 속하며, 추출틀 B에는 속하지 않는 크기 n_a인 표본, 추출틀 A에서 추출되었지만 매칭 이후 추출틀 B에 속하는 크기 n_{ab}인 표본, 추출틀 B로부터 독립적으로 추출된 크기 n_b인 표본이다. 이러한 표본을 기반으로 백신의 포함률을 추정할 수 있다.

이와 같이 세 가지 표본을 모두 사용하여 예방접종 적용률을 추정하고자 한다. 여기서는 Hartley(1962)유형의 설계기반 비추정량(Cochran,

1977)을 이용하여 다음과 같이 주어진다.

$$\hat{R}=\frac{\hat{Y}_a+p\hat{Y}_{ab}+(1-p)\hat{Y}_b}{\hat{N}_a+p\hat{N}_{ab}+(1-p)\hat{N}_b}$$

여기서 $\hat{Y}_a$는 B에 속하지 않는 추출틀 A에서 추출한 표본을 기준으로 특정 백신 또는 백신 시리즈에 대해 최신의 추정된 어린이 수이며, $\hat{Y}_{ab}$은 B에 속하는 추출틀 A에서 추출한 표본을 바탕으로 한 최신의 어린이 수, $\hat{Y}_b$는 추출틀 B의 표본에 기초하여 최신의 추정된 어린이 수이다. $\hat{Y}_{ab}$ 과 $\hat{Y}_b$는 모두 추출틀 B의 최신 아동 수를 추정하는 것이다. 분모에 있는 추정치는 세 표본에 근거한 19개월에서 35개월 사이의 어린이의 추정 수이다. $\hat{N}_{ab}$과 $\hat{N}_b$는 모두 추출틀 B에 있는 어린이의 총수 N_b의 추정치이다. p와 $(1-p)$는 추출틀 B와 관련된 추정치를 결합하는 가중치 요인이다.

한편 m_0의 조건 하에서 $\hat{R}$의 조건부 분산을 최소로 하는 p,m,m_b를 구할 수 있다. $\hat{R}$의 분자와 분모에 있는 총합들과 $\hat{R}$의 분산은 영역 가중치를 이용하여 추정할 수 있는데 $\hat{R}$의 분자에 있는 각 추정치들을 살펴보면 다음과 같다.

$\hat{Y}_a=\frac{M}{m}y_a$이며, 여기서 y_a는 추출틀 B에는 속하지 않고, 추출틀 A에 속하는 n_a명의 어린이들 중에서 최신의 어린이 수이다. 한편 $\hat{Y}_a$는 다음과 같이 다시 표현할 수 있다.

$$\hat{Y}_a=\frac{M}{m}n_a r_a$$

여기서 r_a는 n_a명의 어린이들 중에서 최신의 어린이들의 비율이다.

이와 유사하게 $\hat{Y}_{ab}$ 또한 다음과 같이 표현할 수 있다.

$$\hat{Y}_{ab}=\frac{M}{m}n_{ab}r_{ab}$$

여기서 r_{ab}는 n_{ab}명의 어린이 표본에 기반하여 최신의 추정된 어린이비율이다.

또한 $\hat{Y}_b$는 위에서와 같은 형식으로 표현할 수 있다.

$$\hat{Y}_b = \frac{M}{m} n_b r_b$$

여기서 r_b는 n_b명의 어린이들 중에서 최신의 어린이들의 비율이다.

다음으로 $\hat{R}$의 분모에 있는 합계들의 추정치를 살펴보도록 한다. 추출틀 A에 속하는 추정된 어린들의 총수는 $\hat{N}_a = (M/m)n_a$이며, 추출틀 B에 속하는 추출틀 A에 있는 추정된 어린이의 총수는 $\hat{N}_{ab} = (M/m)n_{ab}$이고, 추출틀 B에 속하는 추정된 어린이의 총수는 $\hat{N}_b = (M/m)n_b$가 된다.

다음으로 $\hat{R}$의 분산을 구해보도록 한다. 먼저 추출틀 A로부터 추출된 어린이의 수 n_a와 추출틀 B에서 추출된 어린이의 수 n_b 들은 반복추출에서 고정된 값이며 이 값들의 기댓값으로 대체될 수 있다는 가정하에서 $\hat{R}$의 조건부 분산을 도출한다.

그러면 $E(n) = m\frac{N}{M} = me$ 이고, $E(n_b) = m_b\frac{N_B}{M_B} = m_b e_b$이다.

또한 $E(n_a) = m\frac{N_a}{M} = m\frac{N_a}{M_a}\frac{M_a}{M} = me_a(1-\alpha)$이며,

$E(n_{ab}) = m\frac{N_{ab}}{M} = m\frac{N_{ab}}{M_{ab}}\frac{M_{ab}}{M} = me_b\alpha$ 이다.

n이 고정이기 때문에 $n_b = n - n_a$이다. 즉, $\hat{N}_b = N - \hat{N}_a$이다. 따라서 $\hat{R}$은 다음과 같이 쓸 수 있다.

$$\hat{R} = \frac{\hat{Y}_a + p\hat{Y}_{ab} + (1-p)\hat{Y}_b}{\hat{N}_a(1-p) + pN + (1-p)\hat{N}_b}$$

그러면 $\hat{N}$과 $\hat{N}_b$를 고정하고 $\hat{R}$의 분산을 다음과 같이 도출할 수 있다.

$$V(\hat{R}) = \frac{1}{N^2}\left[\frac{M^2}{m}R(1-R)\{e_a(1-\alpha)+p^2e_b\alpha\}\frac{M_b^2}{m_b}(1-p)^2e_b\alpha R(1-R)\right]$$

마지막으로 판별 표본크기 $m_0 = m + m_b$에 대해 분산을 최소로 하는 표본크기 m, m_b 및 가중비 p를 구할 수 있다.

$$m = \frac{m_0\alpha\sqrt{e_a}}{\alpha(\sqrt{e_b}-\sqrt{e_a})+\sqrt{e_a}}$$

$$m_b = \frac{m_0\alpha(\sqrt{e_b}-\sqrt{e_a})}{\alpha(\sqrt{e_b}-\sqrt{e_a})+\sqrt{e_a}}$$

$$p = \sqrt{\frac{e_a}{e_b}}$$

3. 이중추출틀 전화조사에서 최적 할당의 문제[4)]

가. 배경

미국의 현대 무작위 전화걸기(RDD) 전화조사는 유선전화 표본과 휴대전화 표본의 두 가지 표본을 사용한다. Wolter, Smith and Blumberg (2010)의 연구는 이러한 이중추출틀 전화조사에 대한 통계적 기초를 제공한다. 여기서는 그들의 연구를 바탕으로 총 조사 자원을 두 표본 추출틀에 할당하기 위한 고려사항과 통계적 방법에 대해 기술한다.

단위당 비용이 적게 들고 사용 이력이 더 길기 때문에, 유선전화 표본은 종종 더 큰 표본이 되고 이 표본의 모든 응답자에 대해 조사 면접을 시도한다. 소규모 휴대전화 표본에 대한 면접 프로토콜은 (1) 응답자 전원에 대한 설문면접을 완료하려고 시도하거나, (2) 응답자의 전화 상태를

4) 본문은 Wolter, Tao, Montgomery, and Smith (2015)를 참고하여 편집한 내용임.

확인하기 위해 간단한 판별 면접을 실시한 후, 전화기를 가진 응답자에 한하여 설문면접을 완료하려고 시도하며, 이 경우 휴대전화 전용(CPO)으로 분류된다. 즉, 판별에서 가정 내에는 유선전화기가 없다고 보고하는 응답자들이다. 판별과정에서 CPO 응답자와 가구 내에 유선전화가 있지만 유선전화로 연결이 안 되는 다른 응답자 모두를 인터뷰하는 것과 같이 다양한 형태의 판별방법이 존재한다.

시간이 지남에 따라 유선 전용(LLO) 모집단(즉, 가정에서 작동하는 유선전화를 보유하고 있지만 휴대전화에 접근할 수 없는 사람)의 규모가 감소함에 따라 조사 통계학자는 휴대전화 표본이 더 큰 표본이 되고 모든 응답자가 인터뷰하는 새로운 설계를 고려할 수 있다(Blumberg and Luke, 2010). 반면에 더 규모가 작은 유선전화 표본에 대한 인터뷰 프로토콜은 판별하거나 모든 응답자가 인터뷰하도록 요구한다. 하지만 여기서는 휴대전화 표본이 일반적으로 더 작은 표본이고, 이 표본에서 응답자들에 대해 전수 또는 판별 프로토콜을 적용하는 것으로 가정한다.

표본 크기가 완료된 사례(즉, 무응답이 없는 경우)를 나타내는 이상적인 가정 하에서 최적의 할당방법을 제시한다. 표본추출 단위(전화번호)와 분석 단위(예: 가구) 사이에는 본질적으로 일대일 관계가 있으며, 유선전화 모집단에는 기본적으로 일대일 관계가 있다. 휴대전화 모집단의 표본추출 단위와 분석 단위 사이의 일대일 관계, 그리고 목표 모집단의 모든 단위는 두 개의 표본추출틀 중 적어도 하나에 포함된다고 가정한다.

이러한 가정들을 고려할 때, 각각의 특정한 분석 단위는 유선, 휴대전화 회선 또는 유선 및 휴대전화 회선과 모두 연결되어 있으며, 최대 한 개의 유선 전화 회선이나 최대 한 개의 휴대전화 회선과 연결된다.

Hartley(1962, 1974), Fuller and Bumeister(1972)를 포함한 이중 추출틀 조사와 관련한 과거 대부분의 문헌은 다양한 표본추출틀에 표본 크기를 할당하는 문제보다는 추정절차에 대해 연구하였다. Skinner and

Rao(1996), Lohr and Rao(2000, 2006). Biemer(1984)와 Lepkowski and Groves(1986)는 특별한 리스트로 보완된 권역 표본처럼 하나의 추출틀이 다른 추출틀의 부분집합일 때 표본할당을 검토하였다.

우선 기본 가정과 기호를 정의하자. U^A는 유선전화 모집단, U^B는 휴대전화 모집단이라 하자. 전체 관심 모집단은 $U = U^A \cup U^B$이다. 일부 단위는 휴대전화와 유선전화 둘 다에 포함될 수 있으며, 다른 한편으로는 유선전화만 소유한 경우의 LLO와 휴대전화만 소유한 경우인 CPO로 구분할 수 있다. 따라서 두 집단의 중복된 부분은 $U^{ab} = U^A \cap U^B$이며, 또한 $U^a = U^A - U^{ab}$, $U^b = U^B - U^{ab}$ 이다. 여기서 U^a는 유선전화(LLO) 영역이며, U^b는 휴대전화(CPO) 영역, U^{ab}는 중복영역이다. 모집단 크기는 $N_A = n(U^A)$이며, $N_B = n(U^B)$, $N_{ab} = n(U^{ab})$, $N_a = n(U^a)$, $N_b = n(U^b)$이며, 여기서 $n(\cdot)$는 크기를 나타내는 함수이다. 한편 중복 집단(중복 사용자)의 비율은 $a = N_{ab}/N_A$이고 $\beta = N_{ab}/N_B$이다.

s_A는 U^A로부터 단순임의 비복원 추출된 표본이며, s_B는 U^B로부터 단순임의 비복원 추출된 표본이고, $n_A = n(s_A)$, $n_B = n(s_B)$로서 완전히 조사가 완료된 표본크기이다. 한편 포함영역 (a, ab, b)는 표본추출당시에는 알 수 없다고 가정한다.

Y_i는 전체 모집단에서 i번째 단위에 대한 관심변수라 하자. 모집단 영역 평균과 분산을 각각 $\overline{Y}_A$, $\overline{Y}_B$, $\overline{Y}_{ab}$, $\overline{Y}_a$, $\overline{Y}_b$이며, S_A^2, S_B^2, S_{ab}^2, S_a^2, S_b^2이다. 여기서의 목적은 전체 모집단 총합 Y를 추정하기 위한 것이다.

나. 전수조사(TAKE-ALL) 프로토콜

전체조사 프로토콜에서는 표본 s_A와 s_B 둘 다에 있는 모든 단위들에 대해 조사 면접을 수행한다. 그러므로 데이터 수집비용은 다음과 같이 근사

적으로 모형화할 수 있다.

$$C_{TA} = c_A n_A + c_B n_B,$$

여기서 c_A는 표본 s_A에서 완료된 면접 당 비용이며, c_B는 표본 s_B에서 완료된 면접당 비용이다.

휴대전화 표본에서 조사 면접의 기대 수는 CPO단위에 대해 $(1-\beta)n_B$ 이며, 이중 사용자에 대해서는 βn_B 이다.

모집단 총합에 대한 비편향 추정량은 다음과 같다(Hartely, 1962).

$$\hat{Y} = \hat{Y}_a + p\hat{Y}_{ab} + q\hat{Y}_{ba} + \hat{Y}_b$$

여기서 p는 혼합 모수이며, $q = 1-p$이며, $\hat{Y}_a = (N_A/n_A)y_A$는 LLO 총합추정량이며, $\hat{Y}_{ab} = (N_A/n_A)y_{ab}$는 유선전화 표본으로 나온 이중 사용자합계의 추정량이며, $\hat{Y}_{ba} = (N_B/n_B)y_{ba}$는 휴대전화 표본으로부터 나온 이중사용자 합계의 추정량, $\hat{Y}_b = (N_B/n_B)y_b$는 CPO총합 추정량이다.

또한 y_a는 s_A와 도메인 U^a의 관측치에 대한 관심변수의 합이고, y_{ab}은 s_A의 관측치에 대한 관심변수의 합이며, y_{ba}는 s_B와 도메인 U^{ab}의 관측치에 대한 관심변수의 합이고, y_b는 s_B와 도메인 U^b의 관측치에 대한 관심변수의 합이다.

고정된 p하에서 $\hat{Y}$의 분산은 다음과 같다.

$$V(\hat{Y}) = N^2\left(\frac{Q_A^2}{n_A} + \frac{Q_B^2}{n_B}\right)$$

여기서 $W_A = N_A/N,\ W_B = N_B/N$ 이며,

$$Q_A^2 = W_A^2\left[(1-\alpha)S_a^2 + \alpha p^2 S_{ab}^2 + \alpha(1-\alpha)(\overline{Y}_a - p\overline{Y}_{ab})^2\right]$$

이고,

$$Q_B^2 = W_B^2\left[(1-\beta)S_b^2+\beta q^2 S_{ab}^2+\beta(1-\beta)(\overline{Y}_b-q\overline{Y}_{ab})^2\right]$$

이다.

이중추출틀에 대한 전체 표본의 전통적인 최적 배정은 다음과 같이 정의할 수 있다(Cochran, 1977).

$$n_{A,opt}=\frac{KQ_A}{\sqrt{c_A}},\quad n_{B,opt}=\frac{KQ_B}{\sqrt{c_B}}$$

여기서 K는 고정된 분산 하에서 비용을 최소화 하거나, 고정된 비용 하에서 분산을 최소화하는 목적에 따라 결정되는 상수이다.

고정된 비용 C_{TA}하에서 분산을 최소로 하는 목적함수는 다음과 같이 주어진다.

$$\min\{V(\hat{Y})\}=\frac{(\sqrt{c_A}Q_A+\sqrt{c_B}Q_B)^2}{C_{TA}}.$$

한편 고정된 분산 V_0하에서 비용을 최소화 하는 목적함수는 다음과 같다.

$$\min\{C_{TA}\}=\frac{(\sqrt{c_A}Q_A+\sqrt{c_B}Q_B)^2}{V_0}.$$

다. 판별조사 프로토콜

판별조사 프로토콜에서 유선전화 표본s_A에 있는 모든 단위에 대해 조사 면접을 실시한다. 휴대전화 표본 s_B에 있는 모든 단위들에 대해서 판별 면접을 실시하고, 다음으로 CPO로서 판별된 단위에 대해 조사면접이 실시된다.

따라서 이러한 프로토콜 하에서 자료 수집에 소요되는 기대비용은 다음과 같다.

$$C_{SC} = c_A n_A + c_B{}' \beta n_B + c_B{}'' (1-\beta) n_B$$
$$= c_A n_A + c_B^* n_B$$

여기서 $c_B{}'$은 표본 s_B에서 완료된 판별 통화 당 비용이며, $c_B{}''$은 표본 s_B에서 판별이 완료된 후 조사면접 당 비용이며, $c_B^* = c_B{}' \beta + c_B{}'' (1-\beta)$이다.

n_A는 유선전화 응답자 중 면접이 완료된 사람들의 수이며, n_B는 휴대전화응답자중 면접이 완료된 사람수(휴대전화 비소유자에 대해 판별통화를 한 사람과 휴대전화 응답자 중 판별조사와 면접조사를 완료한 사람들의 수)이다. 그러므로 조사면접이 완료된 기대 총 인원수는 $n_A + (1-\beta) n_B$이다.

전체 모집단 총합에 대한 비편향 추정량은 다음과 같다.

$$\hat{Y} = \hat{Y}_a + \hat{Y}_b$$

여기서 $\hat{Y}_a = (N_A / n_A) y_A$, $\hat{Y}_b = (N_B / n_B) y_b$, $y_A = y_a + y_{ab}$이다.

한편 추정량의 분산은 다음과 같다.

$$V(\hat{Y}) = N^2 \left(\frac{R_A^2}{n_A} + \frac{R_B^2}{n_B} \right)$$

여기서 $R_A^2 = W_A^2 S_A^2$, $R_B^2 = W_B^2 S_B^2 \left\{ 1 - \beta + \beta(1-\beta) \frac{\overline{Y}^2}{S_b^2} \right\}$이다.

따라서 두 추출틀의 최적 표본크기는 다음과 같다.

$$n_{A,opt} = L R_A / \sqrt{c_A}, \quad n_{B,opt} = L R_B / \sqrt{c_B}$$

여기서 L은 고정된 분산하에서 비용을 최소화하거나, 고정된 비용하에서 분산을 최소화하는 목적에 따라 결정되는 상수이다.

고정된 비용 C_{SC}하에서 분산을 최소로 하는 목적함수는 다음과 같이 주어진다.

$$\min\{V(\hat{Y})\}=\frac{(\sqrt{c_A}R_A+\sqrt{c_B^*}R_B)^2}{C_{SC}}$$

한편 고정된 분산 V_0하에서 비용을 최소화하는 목적함수는 다음과 같다.

$$\min\{C_{SC}\}=\frac{(\sqrt{c_A}R_A+\sqrt{c_B^*}R_B)^2}{V_0}$$

4. 이중추출틀 활용 방안

2019년 현재 등록장애인 규모는 2,618,918명으로 집계되고 있다(통계청, 2020). 등록장애인 명부가 우리나라의 전체 장애인을 포괄하고 있는가? 만일 전체를 포괄하지 못한다면 미등록장애인의 규모는 얼마나 되는지를 장애인실태조사를 통해 파악함으로써 전체 장애인을 포괄할 수 있도록 해야 한다. 이를 위해 기존의 장애인실태조사와 등록장애인 명부를 사용하여 이중추출틀을 활용한 조사 등으로 검토가 가능할 수 있다.

따라서 여기서는 지금까지 이중추출틀에 대한 몇 가지 학술적 검토를 바탕으로 기존의 장애인실태조사의 설계 방향에 대해 제언하고자 한다.

가. 기존의 장애인실태조사

기존의 장애인실태조사는 약 4만 가구 이상을 지역별로 추출하여 장애인 가구 여부를 판별한 후 가구 내 장애인(등록 및 미등록)이 있는 가구를 대상으로 심층조사를 실시하는 방법이다.

앞서 다룬 이중추출틀을 이용한 조사의 경우, 하나의 추출틀은 불완전

하지만 포괄범위가 넓은 추출틀과 이를 보완할 수 있는 소규모 추출틀을 이용하여 전체 규모를 추정하는 방법이다.

따라서 기존의 장애인실태조사는 이러한 측면에서 이중추출틀이기보다는 이중추출방법을 이용한 방법이다.

N개의 모집단으로부터 1단계로 n'개의 대규모 표본을 추출하여 이들의 특성을 파악한다. 즉, 장애여부 및 장애유형 등에 대한 정보를 수집하고, 적격 가구에 대해 심층조사를 실시한다.

모집단은 L개의 층으로 구성되고, 1단계 표본으로부터 장애 유형 등을 층으로 구분하면 h층에 속하는 표본수를 n_h'이라 하자. 2단계 표본에서는 n개의 단위로 층화 추출을 하게 된다. 즉, 장애 유형별로 층을 구분하여 추출한다. 즉, h층의 크기는 n_h'이고 여기서 n_h개를 단순임의 추출하여 이들의 관찰값을 y_{hi}라 하면 $n=\sum_h n_h$이고, $n'=\sum_h n_h'$이다.

이때 $W_h=N_h/N$ 이고, $w_h=n_h/n$이다.

그러면 모집단 평균에 대한 추정량은 다음과 같다.

$$\overline{Y}_{st}=\sum_h w_h\overline{y}_{st}$$

한편 분산은 다음과 같다.

$$V(\overline{y}_{st})=S^2\left(\frac{1}{n'}-\frac{1}{N}\right)+\sum_{h=1}^{L}\frac{W_hS_h^2}{n'}\left(\frac{1}{\lambda_h}-1\right)$$

여기서 $n_h=\lambda_h n_h'$, $0\le\lambda_h\le 1$이다.

이때 n'과 λ_h는 고정된 비용하에서 분산을 최소로 하는 최적값을 구해야 한다. 한편 비용함수는 다음과 같다.

$$C=c'n'+\sum c_h n_h$$

여기서 c'은 단위당 판별 비용이며, c_h는 h층의 단위당 조사비용이다. n_h는 확률변수이므로 기댓값을 구하면 다음과 같다.

$$E(C) = C_0 = c'n' + n'\sum c_h \lambda_h W_h$$

이때 $E(n_h) = E(n_h'\lambda_h) = n' W_h \lambda_h$ 이며, λ_h의 최적값은 다음과 같다.

$$\lambda_h = \sqrt{\frac{c'}{c_h}\frac{S_h^2}{(S^2 - \sum W_h S_h^2)}}$$

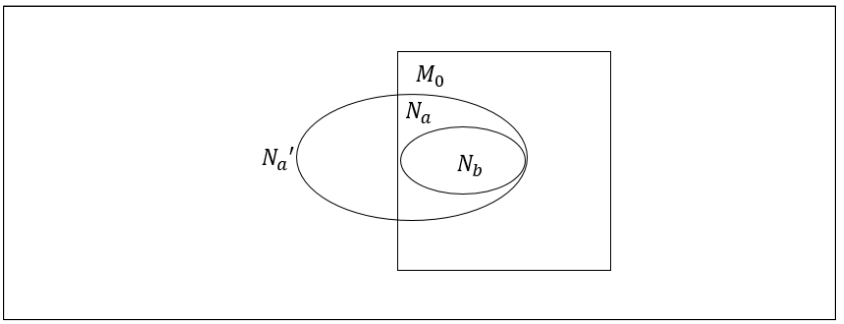
이 값을 비용함수식에 대입하면 n'에 대한 최적값을 구할 수 있다.

기존의 층화 이중추출방법은 대규모 조사가 사전에 실시되고, 이를 바탕으로 심층조사가 실시되기 때문에 가구 및 판별 조사, 심층조사 과정에서 많은 조사비용이 소요되는 문제가 있다. 그러나 등록 및 미등록 장애인을 포괄할 수 있다는 측면에서 타당한 방법이다. 이 방법의 경우 최적의 판별 표본 n'을 얼마로 할 것인지 등에 대해 보다 심층적인 연구가 필요하다.

나. 이중추출틀을 이용한 조사

앞의 방안과는 달리 이 방법은 등록장애인 DB와 등록센서스 기반 조사구를 동시에 활용하는 방법이다. 즉 가구 명부가 추출틀 A이고, 등록 장애인 명부가 추출틀 B이다([그림 3-5] 참조).

〔그림 3-5〕 등록센서스 기반 조사구와 등록장애인 DB 추출틀



자료: 저자 작성.

M_0은 모집단에는 포함되어 있으나, 두 개의 추출틀에는 누락된 부분을 의미한다. N_a와 N_b는 각각의 추출틀 A와 B에 실제 존재하는 단위의 수이며, $N_A = N_a + N_{ab} + N_a'$으로써 추출틀 A에 속한 단위의 총수이며, $N_B = N_b + N_{ab} + N_b'$은 추출틀 B에 속한 단위의 총수를 의미한다. 이때 $N_a' = 0$, $N_b' = 0$으로 가정한다. 따라서 추출틀 A에 대해 $N_A = N_a + N_B$이며, 추출틀 B 에 대해 $N_B = N_b$가 된다.

그러면 관심변수 y의 각 영역별 모평균과 표본평균을 각각 $\overline{Y}_a$, $\overline{Y}_b$, $\overline{y}_a$, $\overline{y}_b$로 정의하면, 모집단 총합에 대한 추정량은 다음과 같이 표현된다.

$$\hat{Y} = N_a\overline{y}_a + N_B(p\overline{y}_B' + q\overline{y}_B'')$$

여기서 $\overline{y}_B'$은 추출틀 A에서 추출된 영역 B에 속한 표본들의 평균이고, $\overline{y}_B''$은 추출틀 B에서 추출된 영역 B에 속한 표본들의 평균이며, $p+q=1$이다. 즉, $\overline{y}_a$는 가구 명부로부터 추출된 장애인 표본들의 평균이고, $\overline{y}_B'$는 가구 명부로부터 추출된 등록 장애인의 평균, $\overline{y}_B''$는 등록 장애인 명부로부터 추출된 표본의 평균을 의미한다.

또한 분산은 다음과 같다.

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\left[S_a^2\left(1-\frac{N_B}{N_A}\right)\right] + N_B S_B^2\left(p^2\frac{N_A}{n_A} + q^2\frac{N_B}{n_B}\right)$$

여기서 S_a^2, S_b^2은 각 영역의 모분산이며, $S_B^2 = S_b^2$이다.

이와 같이 두 개의 추출틀을 이용함으로써 추정의 정도를 높이고, 비용을 절감할 수 있는 장점이 있다. 그러나 추출과정이 복잡하며, 추정량의 최적값 도출이 수치적으로 어렵다는 단점이 있다.

또한 추출틀 A와 추출틀 B 간에서는 서로 보완적인 관계일 경우 가능하며, 대규모 판별조사를 실시할 경우 비용이 적게 소요된다는 측면에서는 유리할 수 있다. 한편 장애인실태조사에서 특정 장애의 경우 전화 조사를 진행하는 데 어려움이 있을 수 있다.

제한된 비용과 자원 하에서 장애인실태조사의 목적을 달성하기 위하여 고려해 볼 수 있는 표본설계방법으로 이중추출틀에 대해 살펴보았다. 등록장애인 DB만으로 포괄할 수 없는 영역을 포괄할 수 있는 방법은 현재와 같이 대규모 표본을 추출하고, 이를 기반으로 판별 조사를 통해 장애인 가구를 판별하여 심층조사를 실시하는 것이 여러 가지 상황에 비추어 볼 때 적합한 방법이라고 볼 수 있다. 한편 이중추출틀은 이용할 추출틀의 부정확성 또는 비포괄성으로 인해 보다 정도가 높은 추출틀을 보완적으로 이용하는 방법이다. 현재 통계청의 가구리스트는 가장 최신 자료이며, 미등록 장애인을 포함하고 있기 때문에 보다 폭넓게 장애인 실태를 파악할 수 있다. 하지만 만일 기존의 표본설계방법을 이용할 경우에는 1단계 표본(n')의 최적값을 산출하여 비용과 시간을 절약할 수 있는 방안을 검토해야 할 것이다.

제3절 통계적 추정 방법

이 절은 장애인 규모에 대해 다른 새로운 방법을 이용하여 추정해 보고, 기존 방법과 새로운 방법의 비교를 통하여 더 정확한 추정이 가능한지 가늠해 보고자 한다.

Capture-Recapture 방법(이하 C-R 방법)은 주로 생태학 분야에서 개체군의 규모를 추정하기 위해 사용하는 방법이다. 최근에는 생태학 분야 이외에도 보건학(epidemiology) 분야에서 특정 질병 총 환자 수, 마약 사용자 수, 도시의 노숙자 인원수 등과 같이 인구 모집단을 추정하는 데 많이 활용되고 있다.

C-R 방법 중 데이터가 2가지인 경우에 활용할 수 있는 여러 추정방법들을 살펴본 후 세부층을 나누어 추정할 경우 적용이 용이한 Lincoln-Petersen method를 사용하여 모집단인 전체 장애인의 수를 추정한다.

Lincoln-Petersen method(이하 L-P 방법)는 오직 두 번의 추출인 포획(Capture)과 재포획(Recapture)이 발생한 경우만을 다루는 가장 대표적이고 간단한 모형이다. '모집단의 크기는 조사기간 동안에 변하지 않아야 하며 포획과 재포획 시 표본을 추출하는 포획확률이 상수로 동일하다'라는 가정이 필요한 방법이다. 이 연구에서의 포획 상황은 장애인 등록 명부에 등록되어 있는지의 여부이며 재포획 상황은 장애인 실태 조사를 통해 파악된 장애인의 명부이다.

1. Capture-Recapture model 가정

사람과 관련된 모집단 크기 추정에 있어서 closed population 가정은 현실적이지 않은 경우가 많고 수집된 데이터(명부)의 완전한 독립은 힘든

경우가 많아 C-R 방법의 적용에 있어 주의를 기울여야 한다. C-R 방법의 적용을 위하여 추후 분석에 있어서 다음의 사항들을 가정하기로 한다.

(1) (closed population) 모집단은 새로운 개체의 유입(출생 또는 이주)과 제거(사망 혹은 이민)에 폐쇄적이다. 즉, 모집단 내부에서의 출생과 사망 또는 외부로부터 새롭게 들어오거나 나가는 이주나 이민이 발생하지 않는다고 가정한다.

(2) 각 개체들이 표본으로 포획될 확률은 동일하다. 각 개개인마다 표본을 추출하는 시기, 행동반응, 그리고 개체들 고유의 차이인 이질성으로 인해 실제로는 만족시키기 어려운 가정이지만 C-R 모형에서는 각 개체들이 표본으로 포획될 확률을 동일하다고 가정한다.

(3) 각 개체들에게 부여된 표식은 조사기간 동안 잃어버리지 않는다. 표식을 잃어버리게 되면 모집단이 폐쇄적이라는 기본 가정을 위반하게 되는 요인 중 하나가 된다. 따라서 조사기간 동안에는 표식을 잃어버리지 않는다고 가정한다.

(4) 연구자는 표식을 정확하게 인식한다. 즉, 이전에 표식을 한 개체에 붙어있는 표식을 연구자가 보지 못한 경우 위배되는 가정으로 연구자가 연구를 주의 깊게 설계하고 집중한다면 감소시키거나 피할 수 있는 문제이다.

(5) 각 개체들은 독립적(independence between individuals)이며, 각 데이터(명부) 역시 서로 독립적(independence between data sources)이어야 한다.

2. 추정 방법

가. Lincoln-Petersen Estimator

Lincoln-Petersen 추정량 (이하 L-P 추정량)은 임의의 표본을 포획하여 표식을 한 후 풀어주고 다시 임의표본을 포획하여 표식이 붙은 동물의 수를 조사함으로써 모집단의 크기를 추정하는 방법이다. 다음의 기호를 가정한다.

N = 모집단 수
m_j = j번째 포획 시도에서 표시된 동물의 수
n_j = j번째 포획 시도에서 포획한 동물의 수

첫 번째 포획 시도에서 n_1 만큼의 표본을 포획하여 표식을 했다고 가정하면, 모집단 중 표식된 표본의 비율 p_1은 다음과 같다. 이때 비율 p_1은 모집단 수를 알 수 없기 때문에 직접 계산할 수 없는 값이다.

$$p_1 = \frac{n_1}{N}$$

두 번째 포획 시도에서 n_2 만큼의 표본을 포획하고, 이들 중 m_2만큼의 표본에 표식이 되어 있었다면, 표식이 된 표본의 비율은 다음과 같다.

$$\hat{p_2} = \frac{m_2}{n_2}$$

첫 번째 시도에서 표식이 된 표본의 비율과 두 번째 시도에서의 비율이 같다고 가정하면, 다음과 같은 식이 성립한다.

$$\frac{m_2}{n_2} \approx \frac{n_1}{N}$$

이 식을 통해 모집단을 추정하면 다음과 같이 추정할 수 있는데 이를 L-P 추정량이라고 한다.

$$\widehat{N}_p = n_1 \frac{n_2}{m_2}$$

L-P 추정량에 대한 분산 추정량과 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$var(\widehat{N}_p) = \frac{n_1 n_2 (n_1 - m_2)(n_2 - m_2)}{m_2^3}$$

$$\widehat{N}_p \pm z_{\alpha/2} var(\widehat{N}_p)$$

나. Chapman Estimator

L-P 추정량은 불편추정량이 아니고 일반적으로 모집단의 크기를 과대추정한다. 따라서 L-P 추정량보다 편의가 작은 수정된 추정량이 등장하는데 이것이 Chapman 추정량이다. 비편향 추정량인 Chapman 추정량은 다음과 같이 모든 항들에 1을 더한 추정량이며,

$$\widetilde{N}_c = \frac{(n_1+1)(n_2+1)}{m_2+1} - 1$$

추정분산과 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$V(\widehat{N}_c) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$$

$$\widehat{N}_c \pm z_{\alpha/2} var(\widehat{N}_c)$$

L-P 추정량에 대한 분산 추정도 chapman 분산 추정량을 쓰는 것이 비편향 성질을 만족한다는 것으로 알려져 있다.

다. Chao's Estimator

Chao 추정량(Chao, 1987, 1989)은 Capture-Recapture model의 가정들 중에서 '각 개체들은 독립적(independence between individuals)이며, 각 데이터(명부)는 서로 독립적(independence between data sources)이어야 한다'라는 가정을 완화하는 추정량이다.

다음과 같이 모수가 2인 혼합된 이항분포를 가정한다.

$$E(f_j) = N\int_0^1 \binom{2}{j} p^j (1-p)^{2-j} f(p)dp,\ j = 0,1,2$$

이때 f_j는 j번째에 나타난 수로 $f_1 = n_1 + n_2 - 2m_2$이고 $f_2 = m_2$이다. 이를 이용해서 자료에 한 번도 나타지 않은 f_0를 추정하는 식은 다음과 같다는 결과가 알려져 있다.

$$\widehat{f_0} = \frac{f_1^2}{4f_2}$$

$n = n_1 + n_2 - m_2$일 때 전체 모집단의 크기를 추정한 Chao 추정량은 다음과 같다.

$$\widehat{N_{Chao}} = n + \frac{f_1^2}{4f_2}$$

추정분산과 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$V(\widehat{N_{chao}}) = \frac{f_1^2}{4f_2}(\frac{f_1}{2f_2}+1)^2$$

$$\widehat{N_{chao}} \pm z_{\alpha/2} var(\widehat{N_{chao}})$$

3. 장애인구 수 추정

C-R 방법을 이용한 장애인구 수 추정을 실시한다. C-R 방법 적용을 위한 장애등록 명부(등록장애인 여부)와 장애인실태조사의 자료 구조는 다음과 같다.

〈표 3-10〉 장애등록 명부 및 장애인실태조사 자료 구조

장애등록 여부 / 장애인실태조사	예	아니오	전체
예	m_2	$n_2 - m_2$	n_2
아니오	$n_1 - m_2$	n_{00}	
전체	n_1		$n_1 + n_2 - m_2 + n_{00}$

자료: 저자 작성.

장애실태조사에서 관측된 등록장애인은 m_2, 장애실태조사에서 관측된 미등록 장애인은 $n_2 - m_2$이고 등록 장애인 중 장애인실태조사 표본이 아닌 장애인은 $n_1 - m_2$이다. n_{00}는 등록 장애인이 아니며 장애실태조사 표본에도 없는 장애 인구수를 뜻하고, 현재 주어진 두 개의 자료에서 관측할 수 없는 부분이다. 따라서 전체 장애 인구수는 $n_1 + n_2 - m_2 + n_{00}$로 계산할 수 있으며 n_{00} 전체 장애 인구수를 알기 위한 추정 대상이다.

두 자료가 서로 독립이라고 가정하였기 때문에 관측 여부에 대한 실패와 성공의 비(odds)는 서로 같다. 따라서 다음과 같은 식이 성립하고, 이를 통해 n_{00}를 추정할 수 있다.

$$\frac{m_2}{n_1 - m_2} \cong \frac{n_2 - m_2}{n_{00}}$$

$$\widehat{n_{00}} = \frac{(n_2 - m_2)(n_1 - m_2)}{m_2}$$

전체 장애 인구수 추정은 $n_1 + n_2 - m_2 + \widehat{n_{00}}$가 되며 식을 정리하면 L-P 추정량과 같아진다.

$$\begin{aligned} n_1 + n_2 - m_2 + \widehat{n_{00}} &= n_1 + n_2 - m_2 + \frac{(n_2 - m_2)(n_1 - m_2)}{m_2} \\ &= n_1 \frac{n_2}{m_2} \end{aligned}$$

Chao 추정량을 이용해 추정하면 n_{00}의 추정값은 다음과 같다.

$$\begin{aligned}\widehat{n_{00}} &= \widehat{f_{00}} \\ &= \frac{f_1^2}{4f_2} \\ &= \frac{(n_1+n_2-2m_2)^2}{4m_2}\end{aligned}$$

따라서 전체 장애 인구수 추정은 $n_1+n_2-m_2+\widehat{n_{00}}$가 되며 식을 정리하면 전체 장애 인구수에 대한 Chao 추정량은 다음과 같다.

$$\begin{aligned}n_1+n_2-m_2+\widehat{n_{00}} &= n_1+n_2-m_2+\frac{(n_1+n_2-2m_2)^2}{4m_2} \\ &= \frac{(n_1+n_2)^2}{4m_2}\end{aligned}$$

이 연구에서는 L-P 추정량을 이용하여 전체 장애 인구수에 대한 추정을 실시하였다. C-R 방법을 이용한 추정 시 표본수를 장애등록을 한 사람의 수로 나누어준 비율($\widehat{p_2}$)을 구한 다음, 이를 이용한 비추정(ratio estimation)을 실시한다. 전체 표본에 대한 비추정 결과는 다음과 같다.

〈표 3-11〉 전체 표본 장애등록여부 현황 및 비가중/가중 장애등록 비율의 역수

(단위: 명)

장애등록여부		표본수	$\widehat{p_2}$	$\hat{p}_{2w}$
예	아니오			
6,397	152	6,549	1.024	1.028

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

여기서 $\widehat{p_2}$는 표본수를 이용하여 구한 비율로 n_2/m_2이며 전체 표본에서는 $\widehat{p_2}$=6,549/6,397 = 1.024 이다. $\hat{p}_{2w}$는 가중치를 이용하여 구한 비율로 $\sum_{i=1}^{n} w_i / \sum_{i=1}^{n} w_i x_i$이고, x_i가 1이면 등록 장애인임을 나타낸다.

가중치는 장애인실태조사의 최종가중치를 사용하였으며 다음과 같은 과정을 통하여 계산된다. 등록 장애인에 대한 가중치와 미등록 장애인에 대한 가중치는 다음의 가구 가중치 wg_{hi}를 기본으로 각각 사후층화를 통해 계산되었다. 여기서 d_{hi}는 설계 가중치이고 r_{hi}는 무응답 조정 가중치이다.

$$wg_{hi} = d_{hi} \times r_{hi},$$

$$d_{hi} = \frac{S_h}{n_h S_{hi}} \times \frac{M_{hi}}{m_{hi}}$$

$$r_{hi} = \frac{m_{hi}}{r_{hi}}$$

n_h : 층 h의 표본조사구 수

S_{hi} : 층 h의 i번째 조사구의 총 가구수

$S_h = \sum_{i=1}^{N_h} S_{hi}$: 층 h에서 총 가구수

M_{hi} : 층 h의 i번째 조사구 내 가구수

m_{hi} : 층 h의 i번째 조사구 내 조사착수 가구수

r_{hi} : 층 h의 i번째 표본조사구 내 조사완료 가구수

장애인에 대한 사후층화는 미등록 장애인의 경우 레이킹비방법(Raking Ratio Method)을 적용하여 설계 가중치와 무응답 조정가중치로 계산된 가구가중치에 대하여 17개 지역(동, 읍·면 고려)으로 나누고, 가구원수 정보를 이용하여 통계청에서 2016년 가구 추계값으로 사후 조정하였다. 등록장애인의 경우는 2016년 12월의 등록장애인 정보를 이용하여 장애유형, 장애등급(중증/경증), 17개 지역별로 사후 조정하여 장애인에 대한 최종 가중치를 산정하였다.

장애 인구수 추정에 대하여 등록장애인수, 장애인실태조사 추정 장애인 수 및 C-R 방법을 이용한 장애 인구수 추정 결과는 〈표 3-12〉에 나타나 있다. 장애인실태조사에서 사용된 장애인구 수 추정식은 다음과 같다.

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

w_{hij} : 각 응답자에 부여된 가중치

y_{hij} : 각 응답결과

L : 층의 수

n_h : 층 h에서의 표본조사구의 수

m_{hi} : 층 h내 i번째 표본조사구의 응답자 수

y_{hij} : 모든 응답자에 대하여 1의 값을 가짐

장애인실태조사 장애인구 수 추정량의 분산과 표준오차는 다음과 같이 계산된다. 먼저 장애인실태조사 장애인구 수 추정량의 분산은 다음과 같다.

$$var(\hat{Y}) = \sum_{h=1}^{L} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (y_{hi.} - \overline{y_{h..}})^2$$

N_h : 층 h의 모집단 크기

$f_h = n_h / N_h$

$y_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$

$\overline{y_{h..}} = \left(\sum_{i=1}^{n_h} y_{hi.} \right) / n_h$

다음 장애인실태조사 장애인 수 추정량에 대한 표준오차는 다음과 같다.

$$s.e(\hat{Y}) = \sqrt{var(\hat{Y})}$$

C-R 추정장애인수($\widehat{N}_p$)와 C-R 추정 장애인수_가중치($\widehat{N_{pw}}$)는 L-P 추정량으로 각각 다음과 같이 계산된다. 이때 n_1은 등록 장애인수 2,511,051이다.

$$\widehat{N}_p = n_1 \times \widehat{p_2}$$
$$\widehat{N_{pw}} = n_1 \times \widehat{p_{2w}}$$

분석에 사용한 가중치는 사후층화가 이루어진 가중치이기 때문에 장애인실태조사의 최종가중치를 사용한 C-R 추정장애인수는 장애인실태조사 추정 장애인수와 같아진다($n_1 = \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi} x_{hi}$).

C-R 추정 장애인수 $\widehat{N}_p$의 분산 추정량은 다음과 같다.

$$var(\widehat{N}_p) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$$

가중치를 사용한 C-R 추정 장애인수 $\widehat{N_{pw}}$의 분산 추정량과 표준오차는 다음과 같다.

$$var(\widehat{N_{pw}}) = n_1^2 var(\widehat{p_{2w}})$$

$$\widehat{p_{2w}} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} x_{hij}}$$

$$var(\widehat{p_{2w}}) = \sum_{h=1}^{L} var_h(\widehat{p_{2w}})$$

$$var_h(\widehat{p_{2w}}) = \frac{n_h(1-f_h)}{n_h - 1}\sum_{i=1}^{n_h}(g_{hi.} - \overline{g_{h..}})^2$$

$$g_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - x_{hij}\widehat{p_{2w}})}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} x_{hij}}$$

$$\overline{g_{h..}} = \left(\sum_{i=1}^{n_h} g_{hi.}\right)/n_h$$

w_{hij} : 각 응답자에 부여된 가중치

y_{hij} : 각 응답결과

L : 층의 수

n_h : 층 h에서의 표본조사구의 수

m_{hi} : 층 h내 i번째 표본조사구의 응답자 수

y_{hij} : 모든 응답자에 대하여 1의 값을 가짐

x_{hij} : 등록 장애인에 대하여 1의 값을 가짐

〈표 3-12〉 장애 인구 수 추정 결과

(단위: 명)

구분	등록장애인수	장애인실태조사 추정장애인수	C-R 추정장애인수	C-R 추정장애인수 _가중치
추정량	2,511,051	2,580,340	2,570,200	2,580,340
표준오차(S.E)		13,471.0	4,889.3	6,442.4

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 3-12〉의 장애인구 수 추정 결과를 보면 장애인실태조사에서의 추정장애인 수와 C-R 방법을 이용한 추정장애인 수가 각각 2,580,340명과 2,570,200명으로 나타나 근소한 차이를 보였다. 그러나 표준오차는 C-R 방법을 이용하여 추정한 경우 4,889.3으로 장애인실태조사(13,471)보다 크게 작아지는 것을 확인하였다.

다음은 전체 표본을 장애유형, 성·연령, 층화변수(대도시, 중소도시, 농어촌)의 기준으로 각각 나누어 장애인구 수 추정을 실시하였다.

〈표 3-13〉은 장애인의 장애유형에 따른 비율 현황을 나타내고 있다. 안면장애의 경우에 비율이 1.077로 가장 컸고, 이는 가중표본수를 이용한 비율도 마찬가지로 나타내고 있다. 자폐성장애 및 간장애의 경우 장애등록을 한 사람과 표본수가 같아 비율이 1로 나타나 장애유형 중에서 가장 작았다.

〈표 3-13〉 장애유형별 장애등록 여부 현황 및 비가중/가중 장애등록 비율의 역수

(단위: 명)

장애유형	장애등록여부		표본수	$\hat{p}_2$	$\hat{p}_{2w}$
	예	아니오			
지체장애	3,157	42	3,199	1.013	1.016
뇌병변장애	623	20	643	1.032	1.041
시각장애	605	18	623	1.030	1.039
청각장애	800	37	837	1.046	1.049
언어장애	53	1	54	1.019	1.007
지적장애	463	9	472	1.019	1.013
자폐성장애	52	0	52	1.000	1.000
정신장애	181	9	190	1.050	1.037
신장장애	194	7	201	1.036	1.043
심장장애	26	1	27	1.038	1.121
호흡기장애	51	2	53	1.039	1.186
간장애	55	0	55	1.000	1.000
안면장애	13	1	14	1.077	1.184
장루요루장애	81	3	84	1.037	1.079
뇌전증	43	2	45	1.047	1.160
전체	6,397	152	6,549		

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 3-14〉는 장애인의 성·연령에 따른 비율 현황이다. 남성과 여성 모두 90대의 비율이 가장 높으며 30대 남성의 비율이 1.005로 가장 낮게 나타났다.

〈표 3-14〉 성 · 연령별 장애등록 여부 현황 및 비가중/가중 장애등록 비율의 역수

(단위: 명)

성 · 연령	장애등록여부		표본수	$\hat{p}_2$	$\hat{p}_{2w}$
	예	아니오			
10대 남자	116	5	121	1.043	1.032
20대 남자	130	1	131	1.008	1.003
30대 남자	184	1	185	1.005	1.007
40대 남자	401	6	407	1.015	1.019
50대 남자	791	9	800	1.011	1.017
60대 남자	832	10	842	1.012	1.012
70대 남자	834	17	851	1.020	1.027
80대 남자	301	7	308	1.023	1.032
90대 남자	19	2	21	1.105	1.094
10대 여자	78	3	81	1.038	1.043
20대 여자	85	1	86	1.012	1.005
30대 여자	91	2	93	1.022	1.021
40대 여자	208	7	215	1.034	1.034
50대 여자	406	10	416	1.025	1.031
60대 여자	604	12	616	1.020	1.023
70대 여자	826	28	854	1.034	1.045
80대 여자	451	27	478	1.060	1.070
90대 여자	40	4	44	1.100	1.115
전체	6,397	152	6,549		

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

〈표 3-15〉는 층화변수에 따른 비율 현황을 나타내고 있다. 층화는 대도시(서울, 부산, 대구, 인천, 광주, 대전, 울산의 동부), 중소도시(경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 세종시의 동부), 농어촌(9개 도 및 세종시 읍·면 지역)의 3개 층으로 구분하였다. 대도시의 비율이 1.032로 가장 높은 반면에 중소도시가 1.012로 가장 낮게 나타났다.

〈표 3-15〉 층화변수별 장애등록 여부 현황 및 비율 및 비가중/가중 장애등록 비율의 역수

(단위: 명)

층화변수	장애등록여부		표본수	$\hat{p}_2$	$\hat{p}_{2w}$
	예	아니오			
대도시	1,912	61	1,973	1.032	1.039
중소도시	1,859	23	1,882	1.012	1.011
농어촌	2,626	68	2,694	1.026	1.050
전체	6,397	152	6,549		

자료: 비공개 자료에 따른 출처 생략.

〈표 3-16〉은 장애인의 장애유형을 기준으로 기존 장애인실태조사에서 추정한 장애인구 수와 C-R 방법을 이용해 새롭게 추정한 장애인구 수에 대한 비교 현황을 나타내고 있다. 기존 장애인실태조사에서 추정한 인구는 2,580,340명이고, C-R 방법을 이용해 새롭게 추정한 장애인구 수는 장애유형별, 성·연령별, 층화변수별로 각각 2,570,200명, 2,568,790명, 2,566,352명으로 기존보다 적은 수의 추정을 하였다. 가중표본수를 사용한 추정에서는 기존과 비슷한 규모의 장애인구를 추정하는 것으로 나타났다.

〈표 3-17〉은 장애인의 성·연령을 기준으로 기존 장애인실태조사에서 추정한 장애인구 수와 C-R 방법을 이용해 새롭게 추정한 장애인구 수에

대한 비교 현황을 나타내고 있다. 기존 장애인실태조사에서 추정한 인구는 2,580,340명이고, C-R 방법을 이용해 새롭게 추정한 장애인구 수는 장애유형별, 성·연령별, 층화변수별로 각각 2,570,200명, 2,568,788명, 2,566,351명으로 기존보다 적은 수의 추정을 하였다. 가중표본수를 사용한 추정에서는 기존과 동일한 규모로 장애 인구를 추정하였다.

〈표 3-18〉은 층화변수를 기준으로 기존 장애인실태조사에서 추정한 장애인구 수와 C-R 방법을 이용해 새롭게 추정한 장애인구 수에 대한 비교 현황을 나타내고 있다. 기존 장애인실태조사에서 추정한 인구는 2,580,340명이고 C-R 방법을 이용해 새롭게 추정한 장애인구 수는 장애유형별, 성·연령별, 층화변수별로 각각 2,570,201명, 2,568,790명, 2,566,351명으로 나타났는데 기존 장애인실태조사보다 더 적은 수의 장애 인구수를 추정하였다. 가중표본수를 사용한 추정에서는 기존 연구와 동일한 규모의 장애 인구를 추정하는 것으로 나타났다.

지금까지 C-R 방법을 이용해 각각 장애유형, 성·연령, 층화변수를 기준으로 새롭게 추정한 장애인구 수와 기존 장애인실태조사를 통한 장애인구 수의 비교 현황에 대해 살펴보았다. 기준 변수로 사용한 장애유형, 성·연령, 층화변수 모두 기존 장애인실태조사에서 추정한 장애인구 수 보다 적은 수의 장애인구 수를 추정하였음을 알 수 있었다. 또한 층화변수를 사용하였을 때의 추정 장애인구 수가 장애유형, 성·연령을 기준으로 추정하였을 때보다 적은 수의 장애인구 수를 추정하였음을 알 수 있었다.

〈표 3-16〉 장애유형 기준 장애인구 수 추정

(단위: 명)

장애유형	등록장애인수	장애인실태조사 추정장애인수	장애유형별 추정장애인수	장애유형별 추정장애인수 _가중치	성 · 연령별 추정장애인수	성 · 연령별 추정장애인수 _가중치	층화변수별 추정장애인수	층화변수별 추정장애인수 _가중치
비등록	0	69,289	0	0	0	0	0	0
지체장애	1,267,174	1,267,174	1,284,555	1,287,645	1,296,322	1,303,278	1,294,947	1,302,234
뇌병변장애	250,456	250,456	258,498	260,696	256,288	257,512	255,984	257,272
시각장애	252,794	252,794	260,323	262,727	258,604	259,852	258,564	259,958
청각장애	271,843	271,843	284,316	285,138	279,493	281,037	277,888	279,624
언어장애	19,409	19,409	19,786	19,574	19,813	19,830	19,822	19,919
지적장애	195,283	195,283	199,079	197,779	199,034	198,751	199,492	200,516
자폐성장애	22,853	22,853	22,853	22,853	23,303	23,200	23,301	23,363
정신장애	100,069	100,069	105,040	103,776	102,067	102,398	102,280	102,772
신장장애	78,750	78,750	81,591	82,119	80,439	80,836	80,537	80,943
심장장애	5,507	5,507	5,719	6,174	5,625	5,651	5,625	5,644
호흡기장애	11,831	11,831	12,295	14,033	12,048	12,107	12,085	12,134
간장애	11,042	11,042	11,042	11,042	11,233	11,278	11,323	11,382
안면장애	2,680	2,680	2,886	3,173	2,738	2,747	2,727	2,734
장루요루장애	14,404	14,404	14,937	15,542	14,698	14,769	14,689	14,741
뇌전증	6,956	6,956	7,280	8,069	7,085	7,095	7,088	7,105
전체	2,511,051	2,580,340	2,570,200	2,580,340	2,568,790	2,580,341	2,566,352	2,580,341

자료: 비공개 자료에 따른 출처 생략.

〈표 3-17〉 성 · 연령 기준 장애인구 수 추정

(단위: 명)

성 · 연령	등록장애인수	장애인실태조사 추정장애인수	장애유형별 추정장애인수	장애유형별 추정장애인수 _가중치	성 · 연령별 추정장애인수	성 · 연령별 추정장애인수 _가중치	층화변수별 추정장애인수	층화변수별 추정장애인수 _가중치
10대 남자	60,144	62,081	61,246	61,127	62,737	62,081	61,369	61,563
20대 남자	63,826	64,025	65,083	64,983	64,317	64,025	65,278	65,561
30대 남자	83,633	84,232	85,595	85,740	84,088	84,232	85,505	85,910
40대 남자	175,125	178,426	179,106	179,540	177,745	178,426	178,974	179,822
50대 남자	337,018	342,879	344,265	345,711	340,852	342,879	344,395	346,034
60대 남자	321,605	325,483	329,006	330,630	325,471	325,483	328,620	330,444
70대 남자	306,844	315,070	314,769	317,076	313,098	315,070	313,641	315,445
80대 남자	95,013	98,016	97,926	98,409	97,222	98,016	97,117	97,828
90대 남자	6,384	6,982	6,617	6,636	7,056	6,982	6,519	6,569
10대 여자	36,714	38,283	37,593	37,595	38,127	38,283	37,566	37,736
20대 여자	39,183	39,382	40,164	40,148	39,644	39,382	40,016	40,189
30대 여자	38,495	39,310	39,581	39,620	39,341	39,310	39,304	39,503
40대 여자	87,240	90,231	89,778	90,003	90,176	90,231	89,256	89,716
50대 여자	165,658	170,812	169,753	170,348	169,739	170,812	169,352	170,227
60대 여자	232,863	238,177	237,965	239,061	237,489	238,177	237,949	239,286
70대 여자	295,875	309,244	302,342	303,677	305,905	309,244	302,388	304,237
80대 여자	149,557	160,007	153,099	153,664	158,510	160,007	152,921	154,016
90대 여자	15,873	17,701	16,312	16,371	17,460	17,701	16,180	16,255
전체	2,511,050	2,580,341	2,570,200	2,580,339	2,568,977	2,580,341	2,566,350	2,580,341

자료: 비공개 자료에 따른 출처 생략.

〈표 3-18〉 층화변수 기준 장애인구 수 추정

(단위: 명)

층화변수	등록장애인수	장애인실태조사 추정장애인수	장애유형별 추정장애인수	장애유형별 추정장애인수 _가중치	성 · 연령별 추정장애인수	성 · 연령별 추정장애인수 _가중치	층화변수별 추정장애인수	층화변수별 추정장애인수 _가중치
대도시	1,002,238	1,041,106	1,026,153	1,030,304	1,024,759	1,029,249	1,034,213	1,041,106
중소도시	1,164,390	1,177,685	1,191,622	1,196,517	1,190,860	1,196,047	1,178,796	1,177,685
농어촌	344,423	361,549	352,426	353,519	353,171	355,044	353,342	361,549
전체	2,511,051	2,580,340	2,570,201	2,580,340	2,568,790	2,580,340	2,566,351	2,580,340

자료: 비공개 자료에 따른 출처 생략.

제4절 소결

이 장에서는 표본설계 효율화 방안을 위한 다각적 접근으로 표본조사구 축소 관련 모의실험 실시, 이중추출틀에 대한 방법론 고찰, 다른 통계적 방법을 사용한 장애인 규모 추정에 대해 살펴보았다.

제1절에서는 2017년 장애인실태조사를 사용하여 표본조사구 축소 비율에 따른 모의실험을 실시하였고 결과는 다음과 같다. 먼저 조사의 공표범위인 권역별 가구수 기준 최대허용오차는 표본조사구 축소 30% 이하에서는 모두 1% 내외의 값을 가졌다. 전국의 경우도 표본조사구 축소 비율과 상관없이 모두 1% 이하로 나타났다. 지역을 세분화하여 시도별 동부·읍면부별 가구수 기준 최대허용오차의 경우 표본조사구 축소 30% 이하에서는 시도별 동부·읍면부별 최대허용오차는 모두 10% 이하의 값을 가졌다. 표본조사구 축소 40% 이상의 경우 세종특별자치시 동부 및 읍면부에서 10%이상이었으나, 그 외 나머지 지역에서는 모두 10% 이하로 나타났다. 다음은 조사구 축소 비율에 따른 조사대상자의 결과이다. 표본조사구를 10% 축소한 경우, 전체 장애인 규모는 6,121명으로 나타났으며 모집단(6,820명) 대비 89.8%를 차지하였다. 즉, 표본조사구를 10% 축소하면 조사대상자 규모도 약 10% 축소된다고 볼 수 있다. 표본조사구를 20% 축소한 경우에는 5,460명(80.1%), 30%의 경우 4,777명(70.0%), 40%의 경우 4,098명(60.1%)이고 50%의 경우 3,416명(50.1%)으로 나타났다(〈표 3-9〉 참조). 모집단 대비 미등록 장애인 비중은 다음과 같다. 표본조사구를 10% 축소한 경우에는 모집단 대비 미등록 장애인의 비중은 89.2%로 나타났다. 이는 표본조사구를 10% 축소하면 미등록 장애인의 규모도 10.2% 축소된다고 볼 수 있다. 표본조사구를 20% 축소한 경우에는 모집단 대비 미등록 장애인의 비중은 80.3%, 30%의 경우 70.1%,

40%의 경우 59.9%이고 50%의 경우 51%로 나타났다([그림 3-1] 참조).

이상의 모의실험 결과를 보듯이 표본조사구를 축소한 비율만큼 조사대상 규모도 비슷한 비율로 축소되는 것을 확인하였다. 추후 조사방법을 변경한다면 조사의 정확도 및 비용을 종합적으로 고려하여 최대허용오차 기준을 결정하기 위한 심층연구를 실시해야 할 것이다.

표본조사구가 축소되면 가구 및 판별조사 대상 가구와 전체 장애인 규모가 축소된다. 전체 장애인 중에서 등록 장애인은 등록장애인 DB를 활용하여 추가 구축할 수 있는 반면에 미등록 장애인은 조사구 조사를 통해서만 구축할 수 있으므로 조사구 축소 비율을 선택 시 신중하게 접근할 필요가 있다. 또한 2017년 한 해의 조사만을 가지고 모의실험을 실시했기 때문에 일반화할 수 없다는 점을 밝혀 둔다.

제2절에서는 이중추출틀에 대한 이론적 고찰 및 해외 사례를 검토하였고, 장애인실태조사에서의 이중추출틀 활용 방안을 모색하였다. 장애인실태조사에서는 이중추출틀로 인구센서스 기반 조사구와 등록장애인 DB를 고려해 볼 수 있다. 2개의 추출틀을 사용하므로 추정의 정도를 높이고 비용을 절감할 수 있다. 그러나 표본추출 과정이 복잡할 뿐만 아니라 추정량에 대한 최적값 도출이 수치적으로 어렵다.

현재는 단일추출틀에서 이중추출방법을 사용하여 대규모 조사인 1차 조사(가구 및 판별조사)를 실시한 다음, 이를 바탕으로 2차 심층조사를 실시하고 있다. 그래서 많은 조사비용이 소요되는 문제가 있으나, 등록 및 미등록 장애인을 모두 포괄할 수 있다. 이 방법의 보완으로 가구 및 판별조사의 최적 표본 규모에 대한 심층연구가 필요하다.

제3절에서는 장애인 규모 추정 시 Capture-Recapture 방법(이하 C-R 방법)을 사용하여 추정해 보았다. C-R 방법으로 장애인구 수를 추정한 결과는 기존(2017년 장애인실태조사에서의 추정 장애인구 수)과 근

소한 차이로 나타났다. 표준오차의 값은 C-R 방법에서 기존보다 작아지는 것을 확인하였다. 또한 장애유형, 성·연령, 층화변수(대도시/중소도시/농어촌) 기준별 장애인 규모를 추정하였다. 장애유형, 성·연령, 층화변수 모두 C-R 방법이 기존보다 장애인 규모를 적게 추정하는 것으로 나타났다. 특히 층화변수를 사용했을 때 C-R 방법에 따른 추정 장애인구 수가 장애유형, 성·연령을 기준으로 추정했을 때보다 적게 추정되었다.

제4장

결론

제 4 장 결론

장애인실태조사 표본설계 효율화 방안으로 다음과 같이 2가지를 검토하였다. 첫 번째는 조사구를 활용하는 방안이고, 두 번째는 조사구와 등록장애인 DB를 병행하는 방안이다. 목표모집단은 모두 장애인이고 각 방안별 특징을 살펴보면 다음과 같다.

1. 조사구를 활용하는 방안(현행 방안)

첫 번째 조사구를 활용하는 방안은 현재 장애인실태조사에서 실시하고 있어서 현행 유지 방안이라고 볼 수 있다.

조사모집단은 통계청의 인구센서스를 바탕으로 작성된 조사구로 단일 추출틀이다. 일반적으로 조사구는 모집단 전수 리스트를 제외하고 통계적 추출 방법으로 확률추출이 가능함과 동시에 대표성 높은 표본추출이 가능한 표본추출틀이라는 점에서 타당하다고 볼 수 있다. 현 상황의 제한된 조사예산 및 인프라(연구인력, 투입 시간 등의 자원) 하에서 적합한 방안으로 볼 수 있다.

표본추출은 이중추출(double sampling)[5] 방법을 이용한다. 장애인에 대한 정보가 사전에 없고 장애인 발현율이 낮기 때문에 1상(first phase)에서 대규모 가구 및 판별조사를 실시하여 장애인을 식별한다. 2상

5) 이중추출 방법은 표본추출틀에서 층화할 때 필요한 보조 변수 정보가 충분히 들어있지 않아서 표본추출틀이 정확하지 않을 때 사용한다. 1차로 대규모의 표본을 추출하여 층화변수로 사용할 수 있으면서도 응답이 간편한 변수들에 대해 조사를 실시한 다음, 이를 근거로 1차 표본단위를 층화하여 각 층에서 일부의 2차 표본을 추출하는 방법이다(손창균, 홍기학, 이기성, 2006, p.29).

에서는 1상에서의 조사를 통해 판별된 장애인을 대상으로 심층조사를 실시한다. 이러한 방법은 2개의 조사를 실시해야 하므로 조사비용이 많이 소요되기는 하지만 장애인을 모두 포괄할 수 있다. 다만, 1상에서의 표본규모 관련 비용과 시간을 절약할 수 있는 최적값 산출에 대한 심층연구가 필요하다고 생각한다.

이 방안의 장점은 기존 방식으로 조사를 실시하므로 장애인 출현율을 산출할 수 있을 뿐만 아니라 이전 조사와의 종적 비교도 가능하므로 통계량의 신뢰성도 상대적으로 높은 편이다. 실사 관리에 있어서도 그동안의 노하우(knowhow)로 돌발 상황 등에 유연하게 대처할 수 있으므로 안정적이라고 볼 수 있다.

다음은 3가지 우려 사항에 대한 내용이다. 첫 번째, 장애인 여부를 판별하기 위해 1차로 조사하는 가구 및 판별조사를 대규모로 실시하기 때문에 조사비용이 많이 소요된다는 점이다. 두 번째, 해가 거듭될수록 조사환경 변화로 인하여 3만 가구 이상을 조사 완료해야 하는 부담감이 있다는 점이다. 한편, 표본추출틀이 가장 최근의 인구센서스 자료를 활용하지만 시점 차이(2017년 장애인실태조사인 경우 2015년 인구센서스 활용)가 있다. 이로 인해 재개발, 건물용도 변경, 신도시 개발 등의 이유로 조사구 변동이 있으므로 조사 진행의 번거로움도 가중될 수 있다. 현재는 관할 동사무소의 협조로 조사가 원활히 수행되고 있는 편이다. 세 번째, 물가상승에 따른 조사원과 지도원의 인건비, 조사수수료 등과 같은 조사제반 비용의 인상분이 예산에 반영되지 않는다면 향후 예산 부족 문제로 이어질 수 있다는 점이다.

이외에 향후 표본설계 고도화를 위하여 등록장애인 DB 활용 방안을 고려해 볼 수 있다. 예를 들면 인접 4개 조사구를 하나의 블록(block)으로 묶으면 7만 5천 개 정도 된다(60가구 기준 조사구는 30만 개 정도임). 이

들 블록에 속하는 등록장애인 규모를 산출할 수 있다면 1차 추출단위(PSU)로 블록을 선정하고 추출된 PSU 내에서 60가구를 조사하여 이 중에서 10여 명의 장애인을 조사하는 것이다. 그러나 등록장애인 DB의 주소와 조사구의 주소에 대한 매칭[6](matching) 작업이 선행되어야 하며, 이는 중요한 과업이 될 것이다. 한편, 인구센서스의 인구 부문 조사항목에는 활동 제약 문항이 있다. 활동 제약이 있는 사람인 경우 장애인일 가능성이 높을 것이므로 조사구 선정 시 활동 제약 정보를 활용해 볼 수 있다. 현재는 표본설계 시 사용할 수 없는 정보이나 추후 가능하게 된다면 효율적인 표본설계를 구현할 수 있다고 생각한다.

2. 조사구와 등록장애인 DB를 병행하는 방안

두 번째 조사구와 등록장애인 DB를 병행하는 방안은 기존 조사 방식에 추가로 등록장애인 DB도 활용하여 조사하는 방안이라고 볼 수 있다. 즉, 가구 및 판별조사의 비중은 줄이고 장애인에 대한 심층조사를 확대하는 방안으로 볼 수 있다.

조사모집단은 인구센서스를 바탕으로 작성된 조사구와 등록장애인 DB로 이중추출틀이다. 이중추출틀을 사용하여 표본추출하는 방법은 2가지를 고려해 볼 수 있다.

---

6) 「2018년 전국다문화가족실태조사」는 집락의 크기가 일정할 수 있도록 표본추출틀을 재구축하여 조사비용을 절감하고 표본추출의 효율성을 높이고자 하였다. 즉, 다문화대상자 리스트를 기반으로 표본설계를 한다면 최종 추출단위를 국적분류 등으로 직접 층화할 수 있으므로 층화효율은 높일 수 있다. 그러나 추출된 표본이 산재해서 현장조사가 어렵고 조사비용이 과다 소요될 것이다. 표본추출틀의 재구축 과정은 다음과 같다. 다문화 대상자 명부에 등록된 도로명 주소와 각종 경계자료(국가기초구역, 인구센서스 통합조사구, 집계구, 도서정보 등)를 공간 연계한다. 이를 바탕으로 최종 조사구 수 기준이 평균 74명 내외가 되도록 행정구역(시군구) 경계 내에서 인구센서스 조사구를 통합하여 새로운 조사구 경계를 작성하였다(최윤정 외, 2019, p.727).

첫 번째 표본추출 방법은 다음과 같다. 등록장애인 DB를 이용하여 행안부의 읍면동 리스트별 장애인 수, 구성비, 장애유형 및 등급과 같은 특성별 규모 등을 파악한다. 조사목적에 맞추어 읍면동 층화, 표본할당을 고려하여 층화다단추출방법으로 표본을 추출한다. 추출 순서는 읍면동 → 조사구(또는 조사구 블록) → 가구이다. 지역표본추출(area sampling)이므로 대표성을 확보할 수 있고 장애인 출현율도 산출 가능하다. 담당지역별로 조사원을 배치할 수 있으며, 담당조사원이 해당 읍면동의 조사구 조사와 등록장애인 조사를 병행하여 조사를 진행할 수 있다. 또한 투입 조사원 수가 상대적으로 많지 않아 조사원 차이에 따른 조사결과의 차이를 줄일 수 있다. 다만, 해당 읍면동으로 제한하면 조사대상자의 중복문제가 발생할 확률이 높아지는 단점이 있으나 사전에 중복리스트를 제외할 수 있을 것이다. 조사비용 측면에서도 조사대상 읍면동을 제한함으로써 면접원의 이동을 최소화하고 교통비, 숙박비 등의 비용을 일부 줄일 수 있다.

두 번째 표본추출 방법은 이중추출틀인 조사구와 등록장애인 DB를 별개로 하여 각각 조사대상을 추출하는 것이다. 즉, 조사구를 기반으로 하여 가구를 추출한다. 그리고 등록장애인 DB에서 읍면동을 선정한 다음, 선정된 읍면동에서 조사대상자(등록장애인)를 추출하는 것이다. 그런데 일반적으로 조사구 조사의 경우에도 읍면동을 선정하는 층화 과정을 거치므로, 결과적으로 읍면동 선정 후 해당 지역의 조사구를 추출하는 절차로 진행된다. 등록장애인 DB의 명부를 대상으로 조사를 진행할 경우에도 읍면동을 우선 선정하고, 해당 읍면동에서 적정 표본을 추출하여 조사를 진행하게 된다. 결국 조사구 추출과 읍면동 추출을 각각 한다는 것은 동일하게 읍면동을 추출한 뒤 대상 조사구 또는 가구를 선정한다는 점에서 유사하다고 볼 수 있다. 단지 읍면동을 동일하게 통일할 것인지 통일하지

않을 것인지의 문제라고 볼 수 있는 것이다. 각각의 조사 설계가 보다 자율적으로 진행된다는 점에서 표본설계의 장점이 있다. 그러나 조사 대상 읍면동이 늘어날 경우 투입해야 할 조사원 수가 늘어나거나 교통비, 숙박비 등이 증가하여 추가 비용이 발생할 수 있다. 투입 조사원 수의 증가는 결국 조사원에 의한 오차를 크게 할 수 있으며, 조사 기간이 길어지게 될 우려가 있다.

결론적으로 보면 이중추출틀을 기반으로 한 표본추출방법은 조사구와 등록장애인 DB 명부 추출 시 읍면동 기준을 다르게 하기보다는 읍면동을 동일하게 하는 것이 효율적이라고 생각한다. 동일 읍면동을 사용한다면 집락들을 구성하고 있는 조사 단위들의 이질적인(heterogeneous) 특성과 집락 간에 대한 동질적인(homogeneous) 특성도 확인할 필요가 있다. 조사 단위들이 이질적일수록 집락 간 변동이 줄어들어서 동일한 조사 비용으로 보다 효율적인 추정이 가능하기 때문이다.

이 방안의 장점은 인구센서스를 바탕으로 작성된 조사구 조사를 실시하므로 적정한 조사구 규모 하에서 장애인 출현율 산출이 가능하다. 또한 등록장애인 DB도 함께 활용하므로 장애 관련 정보 변수(장애유형, 장애등급 등)를 고려한 표본 배분이 가능하여 희귀질환 장애유형도 표본에서 누락되지 않고 구축할 수 있다. 또한 조사구 조사의 규모가 축소되어 가용할 수 있는 조사예산이 확보되므로 추가로 심층조사를 실시할 수 있다.

다음은 우려 사항으로, 첫 번째, 2개의 추출틀 간 기준시점의 차이로 표본 포함 범위(coverage)가 다르게 되는 상황이 발생할 수 있다는 점이다. 두 번째, 등록장애인 DB는 리스트이므로 동일한 지역으로 묶고 집락을 만들어서 관리하는 등과 관련한 표본관리가 어려울 수 있다. 그리고 등록장애인 DB 표본의 경우 등록장애인 DB에서 추출될 확률과 조사구에서 추출될 확률이 중복될 수 있다는 점이 우려된다. 이에 따라 가중치,

신뢰구간 등과 관련한 산출식이 매우 복잡하여 이중추출틀 접근에 의한 추정법의 심층연구가 필요하다고 생각한다. 세 번째, 이중추출틀을 활용한 조사는 기존 조사에 비해 표본추출에서부터 가중치 산출 관련 통계 업무 및 실사관리가 기존 조사에 비해 더 많은 연구인력이 필요하고, 투입시간도 더 많이 소요된다는 점이다. 그렇기 때문에 충분한 인프라 마련이 선행되어야 할 것이다.

한정된 조사예산이므로 조사구와 등록장애인 DB에서의 표본 배분을 검토해야 한다. 그래서 다양한 표본조사구 축소 비율에 따른 최대허용오차 한계를 살펴보고, 축소된 조사구 조사로 인해 가용할 수 있는 조사예산으로 추가 심층조사 가능한 등록장애인 규모도 가늠해 보았다.

먼저 다양한 표본조사구 축소 비율에 따른 모의실험 결과는 다음과 같다. 조사의 공표 범위인 권역별(대도시, 중소도시, 농어촌)에 대한 단순임의추출 가정으로 95% 신뢰수준 하에서 가구수 기준 최대허용오차 한계는 표본조사구 축소 30% 이하의 경우 모두 1% 내외의 값을 가졌다(〈표 3-7〉 참조). 반면에 표본조사구 축소 40% 이상에서는 모두 1% 이상으로 나타났다. 전국의 경우, 표본조사구 축소 비율과 상관없이 모두 1% 이하로 나타났다. 지역을 세분화하여 시도별 동부·읍면부별 가구수 기준 최대허용오차의 경우, 표본조사구 축소 30% 이하에서는 시도별 동부·읍면부별 최대허용오차는 모두 10% 이하의 값을 가졌다. 표본조사구 축소 40% 이상의 경우 세종특별자치시 동부 및 읍면부에서 10%이상이었으나, 그 외 나머지 지역에서는 모두 10% 이하로 나타났다. 한편 조사구 축소 비율에 따른 전체 장애인 규모는 표본조사구를 10% 축소한 경우 6,121명으로 나타났으며 모집단(6,820명) 대비 89.8%를 차지하였다. 즉, 표본조사구를 10% 축소하면 조사대상자 규모도 약 10% 축소된다고 볼 수 있다. 표본조사구를 20% 축소한 경우에는 5,460명(80.1%), 30%의 경우

4,777명(70.0%), 40%의 경우 4,098명(60.1%)이고 50%의 경우 3,416명(50.1%)으로 나타났다(〈표 4-1〉 참조). 모의실험 결과를 종합해 보면 표본조사구를 축소한 비율만큼 조사대상 규모도 비슷한 비율로 축소되는 양상을 보였다. 가구수 기준 최대허용오차 한계 및 조사대상 규모의 결과를 통하여 기존 표본조사구 규모의 30% 이하까지는 축소 가능하다고 볼 수 있다. 단, 조사의 정확도 및 조사비용을 종합적으로 고려한 최대허용오차 기준 마련 및 조사구 조사를 통해서만 가능한 미등록 장애인의 구축 방안 등에 대한 심층연구를 실시해야 할 필요가 있다.

현재 실시 중에 있는 국내 장애인 관련 실태조사 예산 집행을 기반으로 하여, 2017년 장애인실태조사 예산 기준으로 심층조사가 추가 가능한 표본 수를 산출해 보았다. 단, 조사 사례비 수준에 따라 예산 변동이 가능하고, 조사구와 등록장애인 DB에서 추출된 표본의 중복 가능성을 반영해야 하는 등과 같은 여러 가지 제약 조건이 있다. 이러한 제약 조건 하에서 절감된 조사예산에 맞춰 대략적으로 추가 심층조사 가능한 장애인의 규모를 추정해 보았다. 〈표 4-1〉을 보면 예상되는 전체 장애인 수는 표본조사구 축소 10%인 경우 6,821~7,121명, 20%인 경우 7,260~7,660명, 30%인 경우 7,777~8,277명, 40%인 경우 8,098~8,598명, 50%인 경우 8,416~8,916명으로 나타났다. 표본조사구 축소 비율이 증가할수록 2017년 실태조사 응답 완료 기준 6,594명보다 더 많은 장애인을 조사할 수 있다. 그러나 조사구를 활용한 조사의 표본 규모가 축소되면 미등록 장애인을 구축하는 데 어려움이 발생할 수 있다. 이에 따라 충분하지 않은 미등록 장애인의 규모는 대표성 문제를 야기할 수 있고 가중치 산출에도 어려움이 따를 수 있다. 따라서 표본조사구 축소 시 미등록 장애인의 특성을 파악하여 많이 분포되어 있는 곳은 과대표본추출(oversampling)을 고려해 볼 수 있다. 2017년 실태조사 결과로 보면 장애등록을 하지 않

았을 비율이 등록을 한 경우 보다 높게 나타난 집단의 특징은 다음과 같다. 서울, 광역, 세종 및 동부 또는 농어촌 및 읍면부에 거주하며 성별은 여성이고 연령대는 10대 또는 70대 이상이며 최종 학력이 초졸 또는 고졸이고 안면장애를 가지고 있으며, 본인을 포함한 총 가구원 수가 1명 또는 2명이고 본인을 포함한 총 장애인 수가 2명이며 월평균 소득이 없고, 월평균 총 가구소득이 1분위인 경우가 해당되었다.

〈표 4-1〉 조사구 축소 비율별 추가 심층조사 가능한 장애인 규모

(단위: 명)

	2017년 조사 기준	표본조사구 축소				
		10%	20%	30%	40%	50%
조사구 수 (개)	2,001	1,798	1,602	1,400	1,203	1,001
가구수 (가구)	36,200	32,520	28,987	25,328	21,767	18,095
A : 전체 장애인 수 (명)	6,820	6,121	5,460	4,777	4,098	3,416
B : 추가 심층조사 가능한 등록장애인 수 (명)	-	700~ 1,000	1,800~ 2,200	3,000~ 3,500	4,000~ 4,500	5,000~ 5,500
A+B : 예상되는 전체 장애인 수 (명)	-	6,821~ 7,121	7,260~ 7,660	7,777~ 8,277	8,098~ 8,598	8,416~ 8,916

주: 전체 예산 중 직접비를 70%로 가정하였을 때 표본조사구 축소 10%의 경우 조사비용은 7% 절감으로 추정됨. 표본조사구 축소 20%인 경우 조사비용은 14%, 30%인 경우 21%, 40%인 경우 28%, 50%인 경우 35%임.

자료: 보건복지부, 한국보건사회연구원. (2017). 장애인실태조사[데이터파일]. https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출 및 저자 작성.

부가적인 연구로 장애인 규모 추정에 대해 Capture-Recapture 방법(이하 C-R 방법)을 사용하여 추정해 보았다. C-R 방법으로 장애인구 수를 추정한 결과는 기존(2017년 장애인실태조사에서의 추정 장애인구 수)과 근소한 차이로 나타났다. 표준오차의 값은 C-R 방법에서 기존보다 작아지는 것을 확인하였다. 또한 장애유형, 성·연령, 층화변수(대도시/중소도시/농어촌)를 기준으로 각각 구분하여 장애인구 수도 추정해 보았다.

장애유형, 성·연령, 층화변수 모두 C-R 방법이 기존보다 장애인 규모를 적게 추정하는 것으로 나타났다. 특히 층화변수를 사용했을 때 C-R 방법에 따른 추정 장애인구 수가 장애유형, 성·연령을 기준으로 추정했을 때보다 적게 추정되었다.

〈표 4-2〉는 앞에서 살펴본 기존 및 새로운 방안의 내용을 정리하였다. 이 연구는 장애인실태조사의 표본설계 효율화를 위하여 기존 및 새로운 방안에 대해 다각적으로 검토한 기초연구라고 볼 수 있다. 다만 2017년 장애인실태조사 자료 분석에 한정하였고, 조사차수별 분석 및 추이를 살펴볼 수 없었던 제약으로 인하여 통계적 분석 결과는 일반화할 수 없다는 점을 밝혀 둔다.

현재 실시하고 있는 조사방법은 여러 가지 상황에 비추어 볼 때 적합한 방법이라고 볼 수 있다. 그러나 등록장애인 DB 정보의 활용과 관련해서는 계속 제고해야 하고, 실제 조사에 바로 적용하기보다는 심층연구를 통한 신중한 접근이 필요하다고 생각한다.

〈표 4-2〉 조사구를 활용하는 방안 vs. 조사구와 등록장애인 DB를 병행하는 방안

	조사구를 활용하는 방안	조사구와 등록장애인 DB를 병행하는 방안
목표모집단	장애인	장애인
조사모집단 (표본추출틀)	등록센서스 기반 조사구 (단일추출틀)	등록센서스 기반 조사구 및 등록장애인 DB (이중추출틀)
표본추출방법	이중추출방법	층화다단추출방법
장점	• 장애인 출현율 산출 가능 • 이전 조사와의 종적 비교가 가능하여 통계량의 신뢰성이 높은 편 • 실사 관리 측면에서 보면 축적된 경험으로 인한 유연한 상황 대처 가능	• 적정한 조사구 규모 하에서 장애인 출현율 산출 가능 • 조사구 조사 축소로 인한 가용할 수 있는 조사예산 확보로, 추가 심층조사 가능 • 등록장애인 DB의 활용으로 희귀질환 장애유형도 표본에서 누락되지 않고 구축 가능
단점(우려)	• 대규모 가구 및 판별조사로 인한 많은 조사비용 소요 • 해가 거듭될수록 조사환경 변화로 인하여 3만 가구 이상 조사 완료해야 하는 부담감 • 조사 제반 비용의 인상분이 예산에 반영되지 않는다면 향후 예산 부족 발생	• 2개 추출틀 간 기준시점 차이로 표본 포함 범위가 다르게 되는 상황 발생 • 표본관리 어려움 • 기존 조사에 비해 더 많은 연구인력 필요 및 더 많은 투입 시간 소요
보완 방안	• 1상에서 표본 규모 관련 최적값 산출을 위한 심층연구 필요 • 표본설계 고도화를 위한 등록장애인 DB 활용 방안 마련 (표본추출틀 재구축 등)	• 표본조사구 축소 시 미등록 장애인 구축 방안 마련(과대표본추출 등)

자료: 저자 작성.

참고문헌

정기원, 권선진, 계훈방. (1995). **1995년도 장애인실태조사**. 서울: 보건복지부, 한국보건사회연구원.

변용찬, 서동우, 이선우, 김성희, 황주희, 권선진, 계훈방. (2000). **2000년도 장애인실태조사**. 서울: 보건복지부, 한국보건사회연구원.

변용찬, 김성희, 윤상용, 최미영, 계훈방, 권선진, 이선우. (2006). **2005년도 장애인실태조사**. 서울: 보건복지부, 한국보건사회연구원.

변용찬, 김성희, 윤상용, 강민희, 손창균, 최미영, 오혜경. (2009). **2008년도 장애인실태조사**. 서울: 보건복지부, 한국보건사회연구원.

김성희, 변용찬, 손창균, 이연희, 이민경, 이송희, … , & 이선우. (2011). **2011년도 장애인실태조사**. 서울: 보건복지부, 한국보건사회연구원.

김성희, 이연희, 황주희, 오미애, 이민경, 이난희, … , & 이선우. (2014). **2014년도 장애인실태조사**. 세종: 보건복지부, 한국보건사회연구원.

김성희, 이연희, 오욱찬, 황주희, 오미애, 이민경, … , & 이선우. (2018). **2017년도 장애인실태조사**. 세종: 보건복지부, 한국보건사회연구원.

김성희, 이민경, 오미애, 오욱찬, 황주희, 권선진, 오다은. (2019). **2020년 장애인실태조사 사전연구**. 세종: 보건복지부, 한국보건사회연구원.

손창균, 홍기학, 이기성. (2006). **표본추출 및 관리 매뉴얼**. 대전: 통계청, 한국보건사회연구원.

최윤정, 김이선, 선보영, 동제연, 정해숙, 양계민, …, 황정미. (2019). **2018년 전국다문화가족실태조사 연구**. 서울: 한국여성정책연구원.

보건복지부. (2018). **이용자용 통계정보보고서: 장애인실태조사 2018.** http://meta.narastat.kr에서 2020. 11.30. 인출.

보건복지부, 한국보건사회연구원. (2017). **장애인실태조사[데이터파일].** https://data.kihasa.re.kr/microdata에서 2020. 5. 12. 인출.

통계청. (2020). **장애인현황[데이터파일].** https://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01에서 2020.11.12.인출

〈 기타 참고자료 〉

박홍래. (2006). **통계조사론(개정판).** 서울: 영지문화사.

최진식, 이준석, & 남궁평. (1998). **포획 재포획 모형과 선횡단 모형의 결합밀도 추정향에 관한 연구**. 통계연구, 6, 1-17.

Alfons, A., Holzer, J., & Templ, M. (2013). laeken: Estimation of indicators on social exclusion and poverty. R package version 0.4, 4. Retrieved from http://CRAN.R-project.org/package=laeken. 2020. 11. 30.

Arcos, A., Molina, D., Ranalli, M. G., & del Mar Rueda, M. (2015). Frames2: A Package for Estimation in Dual Frame Surveys. R J., 7(1), 52. Retrieved from http://CRAN.R-project.org/package=Frames2. 2020. 11. 30.

Bankier, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81(396), 1074-1079.

Biemer, P. P. (1984). *Methodology for optimal dual frame sample design*. Bureau of the Census.

Blumberg, SJ., Luke, JV. (2010). *Wireless substitution: Early release of estimates from the National Health Interview Survey, January-June 2010*. National Center for Health Statistics.

Blumberg, SJ., Luke, JV. (2013). *Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2012*. National Center for Health Statistics.

Brick, J. M., Judkins, D., & Morganstein, D. (2002). Two-phase list-assisted RDD sampling. *Journal of Official Statistics*, 18(2), 203-215.

Brittain, S., & Böhning, D. (2009). Estimators in capture–recapture studies with two sources. *AStA Advances in Statistical Analysis*, 93(1), 23-47.

CDC. (2012). National Immunization Survey: Data User's Guide for the 2011. Retrieved from http://www.cdc.gov/nchs/nis/data_files.htm 2020. 11. 30.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783-791.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 427-438.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd Ed. New York: Wiley

Deville, J. C. (1993). Estimation de la variance pour les enquêtes en deux phases. *Manuscript*. INSEE, Paris.

Fuller, W. A., & Burmeister, L. F. (1972). Estimation for samples selected from two overlapping frames. In ASA Proceedings of the Social Statistics Sections, pp. 245–249.

Fuller, W. A., Kennedy, W., Schell, D., Sullivan, G. & Park, H. J. (1989). PC CARP. *Ames, IA: Statistical Laboratory, Iowa State University*.

Hartley, H. O. (1962). Multiple frame surveys. *In Proceedings of the social statistics section, American Statistical Association* (Vol. 19, No. 6, pp. 203-206).

Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhya C*, 36(3):99–118.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*: Series A (General), 149(1), 65-82.

Lepkowski, J. M., & Groves, R. M. (1986). A mean squared error model for dual frame, mixed mode survey design. *Journal of the*

American Statistical Association, 81(396), 930-937.

Lohr, S. L., & Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95(449), 271-280.

Lohr, S., & Rao, J. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Lumley, T. (2014). Survey: Analysis of complex survey samples (R package version 3.30). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://CRAN.R-project.org/ package=survey. 2020. 11. 30.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3), 355-375.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika,* 43(3/4), 353-360.

Raj, D., & Chandhok, P. (1998). Sample survey theory. *Narosa.*

Ranalli, M. G., Arcos, A., Rueda, M. and Teodoro, A. (2013). Calibration estimators in dual frames surveys. Statistical Methods & Applications, 25. doi: 10.1007/s10260-015-0336-5.

Ranalli, M. G., Arcos, A., del Mar Rueda, M., & Teodoro, A. (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications*, 25(3), 321-349.

Rao, J. N. K. and Skinner, C. J. (1996). Estimation in dual frame surveys with complex designs. *In Proceedings of the Survey Method Section, Statistical Society of Canada*, pp. 63-68.

Rao, J. N. K., & Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105(492), 1494-1503.

Rojas, H. A. G. (2014). Teaching Sampling: Selection of samples and parameter estimation in finite population. R package version, 3(1). Retrieved from http://CRAN.R-project.org/package=Teaching Sampling. 2020. 11. 30.

Smith P. J., Hoaglin D. C., Battaglia M. P., Khare M., Barker L.E. (2005). *Statistical Methodology of the National Immunization Survey: 1994-2002*. National Center for Health Statistics, Vital and Health Statistics. 2005, 2(138).

Srinath, K. P., Battaglia M. P., and Khare M. (2004). A dual frame sampling design for an RDD survey that screens for a rare population. *In Proceedings of American Statistical Association Section on Survey Research Methods* (pp. 4424-4429).

Skinner, C. J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86(415), 779-784.

Skinner, C. J., & Rao, J. N. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91(433), 349-356.

Srinath, K.P. (2002). Allocation to Strata in Surveys That Screen for Eligible Populations. In *Proceedings of the International Conference on Recent Advances in Survey Sampling*. Ottawa, Canada.

Templ, M. (2014). CRAN task view: Official statistics & survey methodology. Retrieved from http://CRAN.Rproject. org/view= OfficialStatistics. 2020. 11. 30.

Tillé, Y. & Matei, A. (2012). sampling: Survey Sampling, 2012. Retrieved from http://CRAN.R-project.org/package=sampling. 2020. 11. 30.

Wolter, K. (2007). Introduction to variance estimation. *Springer Science & Business Media*.

Wolter, K. M., Smith, P., & Blumberg, S. J. (2010). Statistical foundations of cell-phone surveys. *Survey Methodology*, 36(2), 203-215.

Wolter, K. M., Tao, X., Montgomery, R., & Smith, P. J. (2015). Optimum allocation for a dual-frame telephone survey. Survey methodology, 41(2), 389-401.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, 31(2), 239-243.

부록

〈부표 A-1〉 표본조사구를 20% 축소할 경우 지역별 동부·읍면부별 표본조사구 추출 개수

(단위: 개)

지역	동부	읍면부	전체
서울특별시	218	-	218
부산광역시	102	-	102
대구광역시	64	-	64
인천광역시	90	-	90
광주광역시	45	-	45
대전광역시	64	-	64
울산광역시	13	-	13
세종특별자치시	6	6	12
경기도	218	58	276
강원도	32	46	78
충청북도	32	45	77
충청남도	26	58	84
전라북도	45	13	58
전라남도	38	102	140
경상북도	38	96	134
경상남도	45	70	115
제주도	19	13	32
전체	1,095	507	`1,602

자료: 비공개 자료에 따른 출처 생략.

〈부표 A-2〉 표본조사구를 30% 축소할 경우 지역별 동부·읍면부별 표본조사구 추출 개수

(단위: 개)

지역	동부	읍면부	전체
서울특별시	190	-	190
부산광역시	90	-	90
대구광역시	56	-	56
인천광역시	78	-	78
광주광역시	39	-	39
대전광역시	56	-	56
울산광역시	11	-	11
세종특별자치시	6	6	12
경기도	190	50	240
강원도	28	40	68
충청북도	28	39	67
충청남도	22	50	72
전라북도	39	11	50
전라남도	34	90	124
경상북도	34	84	118
경상남도	39	62	101
제주도	17	11	28
전체	957	443	1,400

자료: 비공개 자료에 따른 출처 생략.

〈부표 A-3〉 표본조사구를 40% 축소할 경우 지역별 동부·읍면부별 표본조사구 추출 개수

(단위: 개)

지역	동부	읍면부	전체
서울특별시	163	-	163
부산광역시	77	-	77
대구광역시	48	-	48
인천광역시	67	-	67
광주광역시	34	-	34
대전광역시	48	-	48
울산광역시	10	-	10
세종특별자치시	5	5	10
경기도	163	43	206
강원도	24	34	58
충청북도	24	34	58
충청남도	19	43	62
전라북도	34	10	44
전라남도	29	77	106
경상북도	29	72	101
경상남도	34	53	87
제주도	14	10	24
전체	822	381	1,203

자료: 비공개 자료에 따른 출처 생략.

〈부표 A-4〉 표본조사구를 50% 축소할 경우 지역별 동부·읍면부별 표본조사구 추출 개수

(단위: 개)

지역	동부	읍면부	전체
서울특별시	136	-	136
부산광역시	64	-	64
대구광역시	40	-	40
인천광역시	56	-	56
광주광역시	28	-	28
대전광역시	40	-	40
울산광역시	8	-	8
세종특별자치시	4	4	8
경기도	136	36	172
강원도	20	29	49
충청북도	20	28	48
충청남도	16	36	52
전라북도	28	8	36
전라남도	24	64	88
경상북도	24	60	84
경상남도	28	44	72
제주도	12	8	20
전체	684	317	1,001

자료: 비공개 자료에 따른 출처 생략.

간행물 회원제 안내

회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「보건사회연구」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

회원 종류

전체 간행물 회원	보건 분야 간행물 회원
120,000원	75,000원
사회 분야 간행물 회원	**정기 간행물 회원**
75,000원	35,000원

가입방법

문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1~5F
  간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- 한국경제서적(총판) 02-737-7498
- 교보문고(광화문점) 1544-1900
- 영풍문고(종로점) 02-399-5600
- Yes24 http://www.yes24.com
- 알라딘 http://www.aladdin.co.kr